Analysis of the Protein-Coding Content of the Sequence of Human Cytomegalovirus Strain AD169

M. S. Chee, A. T. Bankier, S. Beck, R. Bohni, C. M. Brown, R. Cerny, T. Horsnell, C. A. Hutchison III, T. Kouzarides, J. A. Martignetti, E. Preddie, S. C. Satchwell, P. Tomlinson, K. M. Weston, and B. G. Barrell

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Current Topics in Microbiology and Immunology, Vol. 154

1 Introduction

Large-scale sequence analysis of the AD169 strain of human cytomegalovirus (HCMV) began in this laboratory in 1984 when very little was known about the sequence or location of genetic information in the viral genome. At that time sequence analysis was confined to the major immediate-early gene (STENBERG et al. 1984), a region of the Colburn strain that contained CA tracts (JEANG and HAYWARD 1983), the L-S junction region (TAMASHIRO et al. 1984), and what has been termed the transforming region (KOUZARIDES et al. 1983). This chapter is being written in March 1989 when the sequence is complete except for some remaining polishing of certain areas which is still going on (manuscript in preparation). As far as we know there are no major discrepancies in the data which might lead to the sequence changing although of course this cannot be ruled out. We present a preliminary analysis of the HCMV genome and limit ourselves mainly to the potential protein-coding content of over 200 reading frames.

2 Sequence Analysis

The sequence has been determined by M13 shotgun cloning and chain termination sequencing. In this random approach each base is sequenced many times on average so that the consensus produced should be highly accurate. The sequencing strategy involved applying this random procedure to each *Hin*dIII fragment of the viral genome (ORAM et al. 1982). However, the high G + C content caused severe problems as manifested in the many compressions encountered on the sequencing gels. This entailed resequencing many clones substituting dITP or 7-deazaGTP for dGTP in the reactions to minimize the effect. All sequences have been determined on both strands. Detailed accounts of the methods used are published elsewhere (BANKIER et al. 1987; BANKIER and BARRELL 1989). The sequences at the ends of the genome which were not generated in the *Hin*dIII library were obtained from the *Hin*dIII junction fragments C (equivalent to I andQ) and G (equivalent to K and Q) which were sequenced in their entirety, and from a portion of the *Hin*dIII B (K and H) junction fragment from the *Hin*dIII W/H end to the *Eco*RI site 21.2 kb downstream (WESTON and BARRELL 1986) (Fig. 1). Sequences were also obtained across all the *Hin*dIII sites. Double-stranded sequencing on appropriate overlapping cosmid and plasmid clones (FLECKENSTEIN et al. 1982) confirmed that the sequence was contiguous except for an extra 393-bp fragment which was found between *Hin*dIII T and E, and which we have named *Hin*dIII d. The final map in the prototypical orientation of the viral genome with the *Hin*dIII fragments predicted from the sequence is shown in Fig. 1. As the precise ends of the molecule are not known, we have chosen to number the sequence from the start of the direct repeat (DR1) found by TAMASHIRO et al. (1984). By analogy with the "a" sequence of other herpesviruses, this is the closest feature to the end of the genome (MOCARSKI and

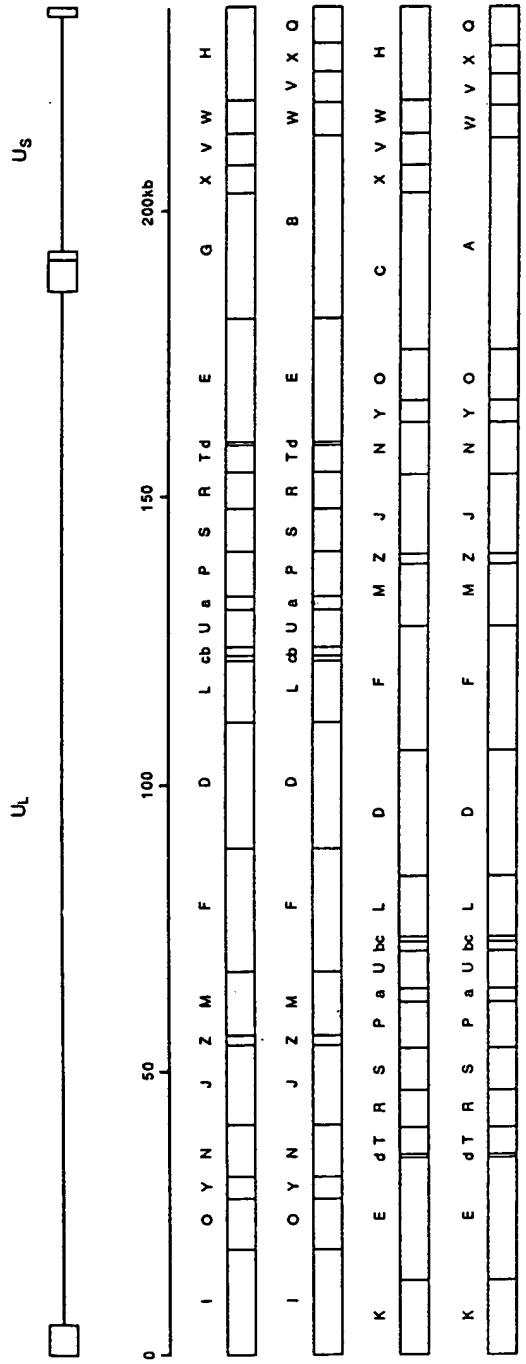


Fig. 1. *Hin*dIII restriction maps of the four HCMV strain AD169 isomers and their relationships to the genome structure (ORAM et al. 1982). The restriction map of the prototype isomer is *topmost* of the four. Individual *Hin*dIII fragments are named alphabetically by size Above the restriction maps a scale is given in kilobase pairs (*kb*). The *uppermost line* shows the genome structure with *UL* (long unique region) and *US* (short unique region) marked; each of these is flanked by their respective repeat sequences shown as blocks

Table 1. A compilation of reading frames of HCMV strain AD169. The orientations, coordinates, and theoretical sizes are tabulated, together with the locations of predicted Kozak consensus ATG codons. For spliced genes exon coordinates represent open reading frame coordinates: donor and acceptor positions are not shown. Lengths are shown in amino acids. References to previous publications in which a *Hin*dIII fragment-based nomenclature is used are as follows: 1 (Weston and Barrell 1986); 2 (Beck and Barrell 1988); 3 (Kouzarides et al. 1988); 4 (Kouzarides et al. 1987b); 5 (Chee et al. 1989b); and 6 (Chee et al. 1989a). References given in the comments section are minimal. Asterisked citations refer to assignments based on other herpesviruses, in particular HSV-1

Frame	Strand	Start	K-ATG	Stop	Length	MW	Old Name	(ref)	Family	Comments
HCMVJ1L	C	3		929	309	33 176		1		Overlaps J11 & J1S
HCMVTRL1		934	970	1 902	311	34 822	HKLF1	1		= HCMVIRL1
HCMVTRL2		1 893		2 237	115	12 324				= HCMVIRL2
HCMVTRL3		3 141	3 192	3 533	114	13 252				= HCMVIRL3. Glycoprotein?
HCMVTRL4	C	3 785		4 435	217	24 929				= HCMVIRL4. ORF in major early transcript (Greenaway and Wilkinson 1987)
HCMVTRL5		4 185	4 266	4 607	114	12 835				= HCMVIRL5
HCMVTRL6	C	5 615	5 947	6 010	111	12 286				= HCMVIRL6
HCMVTRL7	C	6 598	6 843	6 921	82	9 718				= HCMVIRL7
HCMVTRL8		7 227	7 284	7 670	129	14 302				= HCMVIRL8
HCMVTRL9		7 501		7 929	143	15 909				= HCMVIRL9
HCMVTRL10		8 101	8 182	8 694	171	19 035				= HCMVIRL10; D at position 38 is N in IRL10. Glycoprotein
HCMVTRL11		8 648	8 726	9 427	234	26 661			RL11 family	= HCMVIRL11. Glycoprotein
HCMVTRL12		9 431	9 434	10 681	416	47 417			RL11 family	= HCMVIRL12. Glycoprotein
HCMVTRL13		10 778	10 796	11 236	147	15 888				= HCMVIRL13. Glycoprotein exon?
HCMVTRL14		11 140	11 143	11 700	186	21 827			RL11 family	First 35 amino acids identical in IRL14. Glycoprotein exon?
HCMVUL1		11 771	11 810	12 481	224	25 578			RL11 family	Glycoprotein
HCMVUL2	C	12 868	13 047	13 131	60	6 763				Glycoprotein exon?
HCMVUL3	C	13 010	13 324	13 330	105	12 307				
HCMVUL4		13 434	13 464	13 919	152	17 751			RL11 family	Glycoprotein exon?
HCMVUL5		13 986	14 013	14 510	166	18 861			RL11 family	Glycoprotein exon?
HCMVUL6		14 522	14 612	15 463	284	31 447			RL11 family	Glycoprotein exon?
HCMVUL7		15 523	15 526	16 191	222	24 354			RL11 family	Glycoprotein exon?
HCMVUL8		16 198	16 234	16 599	122	13 787			RL11 family	Glycoprotein exon?
HCMVUL9		16 606	16 612	17 295	228	26 889			RL11 family	Glycoprotein
HCMVUL10		17 222		18 199	326	37 366			RL11 family	Glycoprotein exon?
HCMVUL11		18 268	18 295	19 119	275	31 382			RL11 family	Glycoprotein
HCMVUL12	C	19 103	19 321	19 351	73	8 250				Glycoprotein exon?
HCMVUL13		19 143	19 320	20 738	473	54 614				Glycoprotein exon?
HCMVUL14		20 798	20 843	21 871	343	38 567				Glycoprotein

HCMVUL4		13434	13464	13919	152	17751			RL11 family	Glycoprotein exon?
HCMVUL5		13986	14013	14510	166	18861			RL11 family	Glycoprotein exon?
HCMVUL6		14522	14612	15463	284	31447			RL11 family	Glycoprotein exon?
HCMVUL7		15523	15526	16191	222	24354			RL11 family	Glycoprotein exon?
HCMVUL8		16198	16234	16599	122	13787			RL11 family	Glycoprotein exon?
HCMVUL9		16606	16612	17295	228	26889			RL11 family	Glycoprotein
HCMVUL10		17222		18199	326	37366			RL11 family	Glycoprotein exon?
HCMVUL11		18268	18295	19119	275	31382			RL11 family	Glycoprotein
HCMVUL12	C	19103	19321	19351	73	8250				Glycoprotein exon?
HCMVUL13		19143	19320	20738	473	54614				Glycoprotein exon?
HCMVUL14		20798	20843	21871	343	38567				Glycoprotein
HCMVUL15	C	21639		22604	322	35338				
HCMVUL16		22342	22414	23103	230	26148				Glycoprotein
HCMVUL17		23151	23214	23525	104	12672				
HCMVUL18		23631	23637	24740	368	41736	H3O1	2		Glycoprotein homologous to class I HLA (BECK and BARRELL 1988)
HCMVUL19		24701	24740	25033	98	11281				
HCMVUL20		25233	25299	26318	340	38703				Glycoprotein. Homologous to TCR-γ?
HCMVUL21	C	26500	27024	27039	175	19940				
HCMVUL22	C	27263		27646	128	14132				Hydrophobic
HCMVUL23	C	27866		28891	342	39341			US22 family	
HCMVUL24	C	28936	30009	30171	358	40187			US22 family	
HCMVUL25		30030	30057	32024	656	73541			UL25 family	
HCMVUL26	C	32212	32775	32994	188	21156				
HCMVUL27	C	32834	34657	34723	608	69222				
HCMVUL28	C	34757		35893	379	42739			US22 family	
HCMVUL29	C	35926	37005	37092	360	40779			US22 family	
HCMVUL30	C	37138	37500	37533	121	14047				
HCMVUL31		37682		39763	694	76061				
HCMVUL32	C	39850	42993	43050	1048	112689				Large structural phosphoprotein (pp150) (JAHN et al. 1987)
HCMVUL33		43128	43251	44420	390	43806			GCR family	Multiply hydrophobic. Homology to G-protein-coupled receptors
HCMVUL34		44500		46011	504	56185				
HCMVUL35		46042	46093	48012	640	72531			UL25 family	
HCMVUL36EX2	C	48246		49751	408.7	47518	HJLF4	3	US22 family	
HCMVUL36EX1	C	49354	49776	49863	67.3	7483	HJLF3	3	US22 family	
HCMVUL37EX3	C	49913		50842	310	35476	HJLF2	3		IE glycoprotein exon 3
HCMVUL37EX2	C	50893		51015	14.3	1561	HJLF1	3		IE glycoprotein exon 2
HCMVUL38	C	51131	52123	52138	331	36738	HZLF3	3		
HCMVUL37EX1	C	52218	52706	52763	162.7	19116	HZLF2	3		IE glycoprotein exon 1
HCMVUL39		53024		53395	124	13533				
HCMVUL40	C	53216	53878	53893	221	24368				Glycoprotein
HCMVUL41	C	53936		54358	141	16767				
HCMVUL42	C	54384		54854	157	17066				Glycoprotein exon?
HCMVUL43	C	54604	55164	55245	187	20993			US22 family	
HCMVUL44	C	55214	56512	56668	433	46234				Encodes ICP36 protein family (LEACH and MOCARSKI 1989)
HCMVUL45	C	56656		59400	915	101670				Homology to large subunit of ribonucleotide reductase (NIKAS et al. 1986)*

(*Continued*)

Table 1. (*Continued*)

Frame	Strand	Start	K-ATG	Stop	Length	MW	Old Name	(ref)	Family	Comments
HCMVUL46	C	59519	60388	60562	290	33028				Capsid assembly? PERTUISET et al. (1989)*
HCMVUL47		60282	60390	63335	982	109962				
HCMVUL48		62921	63335	70057	2241	253227	HFRF0	4		Virion protein? (BATTERSON et al. 1983*; MCGEOCH et al. 1988a)*
HCMVUL49	C	70403	72112	72334	570	63852	HFLF5	4		
HCMVUL50	C	72072	73262	73283	397	42902	HFLF4	4		Glycoprotein?
HCMVUL51	C	73287	73757	73910	157	16968	HFLF3	4		
HCMVUL52		73748	73796	75799	668	74122	HFRF1	4		
HCMVUL53		75789	75795	76922	376	42314	HFRF2	4		
HCMVUL54	C	76906	80631	80655	1242	137104	HFLF2	4		DNA Polymerase (KOUZARIDES et al. 1987a)
HCMVUL55	C	80775	83492	83654	906	102005	HFLF1	4		gB (CRANAGE et al. 1986)
HCMVUL56	C	83458	86007	86019	850	95870	HFLF0	4		
HCMVUL57	C	86577	90281	90326	1235	133880				Major DNA-binding protein (ANDERS and GIBSON 1988)
HCMVUL58		90864		91235	124	14418				
HCMVUL59	C	91205	91573	91597	123	13945				
HCMVUL60	C	92336		92815	160	18241				
HCMVUL61	C	92847		94139	431	44310				
HCMVUL62	C	94114		94764	217	23686				
HCMVUL63		95331		95717	129	14792				
HCMVUL64	C	95904		96203	100	11245				
HCMVUL65		96315		96620	102	11525				Segments in frame with 67-kDa phosphoprotein sequence of DAVIS and HUANG (1985)
HCMVUL66	C	96475		96816	114	13921				
HCMVUL67	C	97098	97436	97451	113	13218				Glycoprotein exon?
HCMVUL68	C	97750	98079	98100	110	12728				
HCMVUL69	C	98202	100433	100532	744	82679				Transactivator? (MCGEOCH et al. 1988a)*
HCMVUL70	C	100536		103721	1062	120928				DNA replication? (MCGEOCH et al. 1988b)*
HCMVUL71		103239		104471	411	45728				
HCMVUL72	C	104558	105721	105751	388	43576				dUTPase? (PRESTON and FISHER 1984)*
HCMVUL73		105629	105737	106150	138	14868				Glycoprotein
HCMVUL74	C	106128	107525	107585	466	54236				Glycoprotein exon?
HCMVUL75	C	107904	110132	110153	743	84453				gH (CRANAGE et al. 1988)
HCMVUL76		110324	110327	111301	325	36070				

HCMVUL69	C	98 202	100 433	100 532	744	82 679				Transactivator? (McGeoch et al. 1988a)*
HCMVUL70	C	100 536		103 721	1062	120 928				DNA replication? (McGeoch et al. 1988b)*
HCMVUL71		103 239		104 471	411	45 728				
HCMVUL72	C	104 558	105 721	105 751	388	43 576				dUTPase? (Preston and Fisher 1984)*
HCMVUL73		105 629	105 737	106 150	138	14 868				Glycoprotein
HCMVUL74	C	106 128	107 525	107 585	466	54 236				Glycoprotein exon?
HCMVUL75	C	107 904	110 132	110 153	743	84 453				gH (Cranage et al. 1988)
HCMVUL76		110 324	110 327	111 301	325	36 070				
HCMVUL77		110 787	110 907	112 832	642	71 188				Virion protein? (Addison et al. 1984*; McGeoch et al. 1988a)*
HCMVUL78		112 864	112 924	114 216	431	47 358				
HCMVUL79	C	114 277	115 161	115 779	295	33 846				
HCMVUL80		115 084	115 198	117 321	708	73 853				Assembly protein read from internal start (Robson and Gibson 1989)
HCMVUL81	C	117 311		117 658	116	12 796				
HCMVUL82	C	117 489	119 165	119 189	559	61 950			UL82 family	pp71 (Ruger et al. 1987)
HCMVUL83	C	119 355	121 037	121 094	561	62 900			UL82 family	pp65 (Ruger et al. 1987)
HCMVUL84	C	121 312	123 069	123 306	586	65 430				
HCMVUL85	C	123 104	124 021	124 090	306	34 596				
HCMVUL86	C	124 186	128 295	128 415	1370	153 875	HaLF1	5		Major capsid protein (Chee et al. 1989b)
HCMVUL87		128 265	128 355	131 177	941	104 805				
HCMVUL88		131 144	131 177	132 463	429	47 691				
HCMVUL89EX2	C	132 466		133 629	378	42 776				Conserved herpesvirus spliced gene (Costa et al. 1985)*
HCMVUL90	C	133 639	133 836	133 920	66	7 445				
HCMVUL91		133 784	133 835	134 167	111	12 028				
HCMVUL92		134 020	134 140	134 742	201	22 512				
HCMVUL93		134 693	134 711	136 492	594	68 464				
HCMVUL94		136 008	136 353	137 387	345	38 382				
HCMVUL89EX1	C	137 382	138 389	138 803	296	34 323				Conserved herpesvirus spliced gene (Costa et al. 1985)*
HCMVUL95		138 352	138 388	139 980	531	57 214				
HCMVUL96		139 821	140 016	140 360	115	13 108				
HCMVUL97		140 373	140 484	142 604	707	78 234	HSRF3	6		Phosphotransferase? (Chee et al. 1989a)
HCMVUL98		142 626	142 701	144 452	584	65 273				DNase (McGeoch et al. 1986)*
HCMVUL99		144 311	144 392	144 961	190	20 924				Phosphoprotein pp28 (Meyer et al. 1988)
HCMVUL100	C	145 229	146 344	146 413	372	42 862				Multiply hydrophobic
HCMVUL101		146 353		146 697	115	12 184				DNA replication? Position only (McGeoch et al. 1988b)*
HCMVUL102		146 747		149 140	798	85 615				DNA replication? Position only (McGeoch et al. 1988b)*
HCMVUL103	C	149 311	150 057	150 108	249	28 637				
HCMVUL104	C	150 008	152 098	152 167	697	78 508				Virion protein? (Weller et al. 1983*; McGeoch et al. 1988a)*

(*Continued*)

Table 1. (*Continued*)

Frame	Strand	Start	K-ATG	Stop	Length	MW	Old Name (ref)	Family	Comments
HCMVUL105		151 806	151 926	154 793	956	106 501			Helicase (MARTIGNETTI 1987; CRUTE et al. 1989)*
HCMVUL106	C	154 950	155 324	155 330	125	14 500			
HCMVUL107	C	155 420		155 869	150	17 374			
HCMVUL108		156 016		156 384	123	14 501			
HCMVUL109	C	157 517	157 810	157 816	98	11 709			
HCMVUL110	C	157 896		158 276	127	14 224			
HCMVUL111	C	159 479		159 799	107	11 565			
HCMVUL111A		159 615	159 678	159 911	78	8 582			ORF in transforming region (RAZZAQUE et al. 1988)
HCMVUL112		160 484	160 589	161 392	252.3	26 415			Common N-terminus of four phosphoproteins (WRIGHT et al. 1988)
HCMVUL113		161 301		162 797	499	51 105			Probably spliced to UL112; internal splicing? (WRIGHT et al. 1988)
HCMVUL114	C	162 973	163 722	163 758	250	28 354			Uracil-DNA glycosylase (WORRAD and CARAIDONNA 1988)*
HCMVUL115	C	163 697		164 614	306	34 110			
HCMVUL116	C	164 533		165 564	344	37 519			Glycoprotein exon?
HCMVUL117	C	165 474	166 745	166 757	424	45 464			
HCMVUL118	C	166 861		167 487	209	24 599			Glycoprotein exon?
HCMVUL119	C	167 558	167 983	168 037	142	14 729			Glycoprotein exon?
HCMVUL120	C	168 041	168 643	168 700	201	22 768			Glycoprotein
HCMVUL121	C	168 697	169 236	169 269	180	20 138			Glycoprotein
HCMVUL122	C	169 367		170 878	494.7	51 084			IE2A. Spliced to IE1 EX4. Also KATG at 170599 (STENBERG et al. 1985)
HCMVUL123EX4	C	171 009		172 274	405.7	45 622			IE1 gene exon 4 (STENBERG et al. 1984; AKRIGG et al. 1985)
HCMVUL123EX3	C	172 301		172 654	61.7	6 865			IE1 gene exon 3 (STENBERG et al. 1984; AKRIGG et al. 1985)
HCMVUL123EX2		172 659	172 765	172 873	23.7	2 658			IE1 gene exon 2 (first coding exon) (STENBERG et al. 1984; AKRIGG et al. 1985)
HCMVUL124		172 783	172 798	173 253	152	15 887			Glycoprotein
HCMVUL125	C	173 114		173 419	102	11 000			
HCMVUL126	C	173 508		173 909	134	15 910			

[illegible]	C			[illegible]	[illegible]	45 022				IE1 gene exon 4 (STENBERG et al. 1984; AKRIGG et al. 1985)
HCMVUL123EX3	C	172 301		172 654	61.7	6 865				IE1 gene exon 3 (STENBERG et al. 1984; AKRIGG et al. 1985)
HCMVUL123EX2		172 659	172 765	172 873	23.7	2 658				IE1 gene exon 2 (first coding exon) (STENBERG et al. 1984; AKRIGG et al. 1985)
HCMVUL124		172 783	172 798	173 253	152	15 887				Glycoprotein
HCMVUL125	C	173 114		173 419	102	11 000				
HCMVUL126	C	173 508		173 909	134	15 910				
HCMVUL127		174 453	174 495	174 887	131	15 248				Glycoprotein exon?
HCMVUL128	C	174 868		175 284	139	16 036				Glycoprotein exon?
HCMVUL129	C	175 357		175 704	116	13 288				
HCMVUL130	C	175 665	176 306	176 438	214	24 653				Glycoprotein
HCMVUL131	C	176 644	176 871	177 042	76	8 243				First 35 amino acids identical in TRL14
HCMVUL132	C	176 934	177 743	177 845	270	29 973				= HCMVTRL13. Glycoprotein exon?
HCMVIRL14	C	177 776	178 324	178 327	183	20 750				
HCMVIRL13	C	178 231	178 671	178 689	147	15 888				
HCMVIRL12	C	178 786	180 033	180 036	416	47 417			RL11 family	= HCMVTRL12. Glycoprotein
HCMVIRL11	C	180 040	180 741	180 819	234	26 661			RL11 family	= HCMVTRL11. Glycoprotein
HCMVIRL10	C	180 773	181 285	181 366	171	19 034				= HCMVTRL10; N at position 38 is D in TRL10. Glycoprotein
HCMVIRL9	C	181 538		181 966	143	15 909				= HCMVTRL9
HCMVIRL8	C	181 797	182 183	182 240	129	14 302				= HCMVTRL8
HCMVIRL7		182 546	182 624	182 869	82	9718				= HCMVTRL7
HCMVIRL6		183 457	183 520	183 852	111	12 286				= HCMVTRL6. Glycoprotein exon?
HCMVIRL5	C	184 860	185 201	185 282	114	12 835				= HCMVTRL5
HCMVIRL4		185 032		185 682	217	24 929				= HCMVTRL4. ORF in major early transcript (GREENAWAY and WILKINSON 1987)
HCMVIRL3	C	185 934	186 275	186 326	114	13 252				= HCMVTRL3. Glycoprotein?
HCMVIRL2	C	187 230		187 574	115	12 324				= HCMVTRL2
HCMVIRL1	C	187 565	188 497	188 533	311	34 822	HKLF1	1		= HCMVTRL1
HCMVJ11		188 538		189 560	341	36 544				Positions 1 to 309 overlap J1L; 118 to 341 overlap J1S
HCMVIRS1		189 702	189 765	192 302	846	91 050	HQRF1	1	US22 family	V at position 190 is L in TRS1. Sequences diverge after position 549
HCMVUS1	C	192 332		192 967	212	23 481	HQLF3	1	US1 family	
HCMVUS2	C	193 119	193 715	193 850	199	23 112	HQLF2	1	US2 family	Glycoprotein
HCMVUS3	C	194 133	194 690	194 924	186	21 575	HQLF1	1	US2 family	Spliced IE glycoprotein (WESTON 1988)
HCMVUS4		194 832		195 188	119	13 089				
HCMVUS5		195 203	195 230	195 607	126	14 451				
HCMVUS6	C	195 403	195 951	195 975	183	20 640	HXLF6	1	US6 family	Glycoprotein
HCMVUS7	C	196 377	197 051	197 069	225	26 271	HXLF5	1	US6 family	Glycoprotein

(Continued)

Table 1. (*Continued*)

Frame	Strand	Start	K-ATG	Stop	Length	MW	Old Name	(ref)	Family	Comments
HCMVUS8	C	197256	197936	197960	227	26634	HXLF4	1	US6 family	Glycoprotein
HCMVUS9	C	197954	198694	198772	247	28054	HXLF3	1	US6 family	Glycoprotein
HCMVUS10	C	199083	199637	199646	185	20772	HXLF2	1	US6 family	Glycoprotein
HCMVUS11	C	199716	200360	200366	215	25265	HXLF1	1	US6 family	Glycoprotein
HCMVUS12	C	200549	201391	201562	281	32470	HVLF6	1	US12 family	Multiply hydrophobic
HCMVUS13	C	201474	202256	202307	261	29461	HVLF5	1	US12 family	Multiply hydrophobic
HCMVUS14	C	202328	203257	203311	310	34198	HVLF4	1	US12 family	Multiply hydrophobic
HCMVUS15	C	203305		204756	484	53049	HVLF3	1	US12 family	Multiply hydrophobic
HCMVUS16	C	204153	205079	205091	309	34718	HVLF2	1	US12 family	Multiply hydrophobic
HCMVUS17	C	205227	206105	206144	293	31910	HVLF1	1	US12 family	Multiply hydrophobic
HCMVUS18	C	206376	207197	207266	274	30195	HWLF5	1	US12 family	Multiply hydrophobic
HCMVUS19	C	207338	208057	208132	240	26424	HWLF4	1	US12 family	Multiply hydrophobic
HCMVUS20	C	208107		209177	357	39890	HWLF3	1	US12 family	Multiply hydrophobic
HCMVUS21	C	208978	209694	209793	239	26586	HWLF2	1	US12 family	Multiply hydrophobic
HCMVUS22	C	209874		211652	593	66971	HWLF1	1	US22 family	Early nuclear protein (MOCARSKI et al. 1988)
HCMVUS23	C	211717	213492	213510	592	68886	HHLF7	1	US22 family	
HCMVUS24	C	213591	215090	215105	500	57928	HHLF6	1	US22 family	
HCMVUS25		215097		215633	179	19655				
HCMVUS26	C	215730	217536	217574	603	70022	HHLF5	1	US22 family	
HCMVUS27		217859	217904	218989	362	41996	HHRF2	1	GCR family	Multiply hydrophobic. Homology to G-protein-coupled receptors
HCMVUS28		219083	219200	220168	323	37189	HHRF3	1	GCR family	Multiply hydrophobic. Homology to G-protein-coupled receptors
HCMVUS29		220420	220426	221811	462	51068	HHRF4	1		
HCMVUS30		221537	221618	222664	349	39115	HHRF5	1		
HCMVUS31		222674		223264	197	22936	HHRF6	1	US1 family	
HCMVUS32		223325	223385	223933	183	22058	HHRF7	1	US1 family	
HCMVUS33	C	224075		224485	137	15775	HHLF3	1		
HCMVUS34		224408	224480	224968	163	17767	HHRF8	1		Glycoprotein exon?
HCMVUS35	C	225212		225538	109	12966	HHLF2	1		
HCMVUS36	C	225429		225758	110	12352				
HCMVTRS1	C	226115	228478	228541	788	83983	HHLF1	1	US22 family	L at position 190 is V in IRS1. Sequences diverge after position 549
HCMVJ1S	C	228683		229354	224	23797				Overlaps J1L & J1I

HCMVUS31		[illegible]		[illegible]	[illegible]	22 936	HHRF6	1	US1 family
HCMVUS32	C	223 325	223 385	223 933	183	22 058	HHRF7	1	US1 family
HCMVUS33		224 075		224 485	137	15 775	HHLF3	1	
HCMVUS34	C	224 408	224 480	224 968	163	17 767	HHRF8	1	Glycoprotein exon?
HCMVUS35	C	225 212		225 538	109	12 966	HHLF2	1	
HCMVUS36	C	225 429		225 758	110	12 352			
HCMVTRS1		226 115	228 478	228 541	788	83 983	HHLF1	1	US22 family; L at position 190 is V in IRS1. Sequences diverge after position 549
HCMVJ1S	C	228 683		229 354	224	23 797			Overlaps J1L & J1I

ROIZMAN 1982; TAMASHIRO et al. 1984; SPAETE and MOCARSKI 1985b). Our sequence is numbered from base 2352 of TAMASHIRO et al. (1984) but reading backward on the complementary strand. It contains a single copy of a DR1-flanked 578-bp sequence at each end and at the junction of the internal repeats. The sequence we have determined consists of 229 354 base pairs. The long unique region (*UL*) is 166 972 bp and the surrounding repeats (IRL and TRL) are 11 247 bp each. The short unique region (*US*) is 35 418 bp and is flanked by 2524-bp repeats (*IRS* and *TRS*). In the sizes given above, *IRL* and *IRS* are considered as overlapping by one copy of the DR1-flanked repeat unit. The long repeats are identical except for two base changes: a C at position 5288 and a G at position 8293 are both substituted by As in the equivalent *IRL* positions. The former change does not affect any predicted coding sequences, while the latter affects *TRL/IRL10* (Table 1). Two differences were also found in the short repeats: in *IRS*, an A at position 189 887 and a G at position 190 332 are substituted by C and T respectively in *TRS*. The former difference is silent while the latter changes a valine residue in HCMV-IRS1 to a leucine in HCMV-TRS1.

3 Prediction of Reading Frames

Very little of the genome has been mapped in terms of its transcription or its expression. In order to analyze the protein-coding content of the sequence we need to define the criteria for the selection of the reading frames we think are most likely to be coding. A description of the procedures we have applied is given below.

3.1 Criteria for Selection

Analysis of other herpesvirus genomes shows that in most regions the reading frame that is coding is the longest and that such reading frames are arranged end to end on either strand with very little noncoding sequence in between. Very few overlapping genes have been found although there are sometimes small overlaps at the beginnings and ends of genes. Thus the strategy we have adopted has been to screen the sequence for reading frames that are over a certain length and then to filter out any smaller frames that overlap larger ones by a certain amount. The cutoffs that we have chosen are a minimum length of 300 bp (i.e., a coding potential of 100 amino acids) and a maximum allowable overlap of a larger reading frame of 60%. This latter figure allows for the fact that a reading frame may be open upstream of the actual initiation codon and that this may lie under the preceding gene. There are 778 reading frames over 300 bp of which 581 are screened out on the grounds that they are overlapped extensively by larger frames, leaving 197 candidate protein-coding genes. The sequence is then examined for reading frames of less than 300 bp that may lie in the gaps that are left. Likely frames are selected by experience using criteria such as logical combinations of potential transcription signals with the reading

frame and any potential translational start; homology to other reading frames or known genes; and the presence of protein structural or functional motifs in the amino acid sequence. Codon bias can also be used as described below. The whole procedure will not work where genes are spliced and the exons are small. In those regions of the genome where the genes are highly spliced or in regions which are noncoding, small background noncoding reading frames will have been included which would otherwise have been screened out if larger coding reading frames were present. We think that this is particularly true in and bordering the repeat sequences and in certain regions of the *Hin*dIII D and E fragments. In a few cases we have substituted a smaller frame for a larger overlapping frame where we have found compelling reasons to choose the former.

3.2 Codon Bias

Patterns of codon usage that could conceivably be generated only through the genetic code are, in the absence of any other criteria, the best indication that a sequence is coding for protein. The high G + C content of HCMV (57.2%) leads to an accumulation of G and C in the third, degenerative, position of the codons. This is because in an average amino acid sequence the excess G and C cannot be accommodated in the first and second positions without biasing the sequence to amino acids encoded by GC-rich codons. Figure 2 shows a G + C plot across the entire sequence. As can be seen there is considerable variation in the G + C content across the genome, particularly in the repeat areas, the regions bordering the repeats, and the *Hin*dIII D fragment. Because of this variability we have not yet been able to find a single formula that we could apply equally to all areas of the genome to justify further our selection of reading frames on the basis of size and position. However, codon bias does serve as a useful check in those areas with a high G + C content.

3.3 HCMV Map

The preliminary map of 208 reading frames deduced from the sequence using the criteria discussed above is shown in Fig. 3. Details are given in the figure legend of individual frames that we have omitted from the original set of 197 (Sect. 3.1) and the criteria for inclusion of replacement frames. Although some of the frames shown are unlikely to be coding (for example, UL126 which overlaps the (noncoding) exon 1 of the major immediate-early gene and part of the enhancer) we preferred to include all frames meeting our minimal criteria unless a more plausible alternative candidate could be identified.

er reading frames or
ctional motifs in the
ed below. The whole
is are small. In those
in regions which are
. have been included
. reading frames were
the repeat sequences
a few cases we have
here we have found

ed only through the
est indication that a
MV (57.2%) leads to
n of the codons. This
G and C cannot be
sing the sequence to
+ C plot across the
in the G + C content
gions bordering the
we have not yet been
reas of the genome to
of size and position.
as with a high G + C

e sequence using the
1 the figure legend of
97 (Sect. 3.1) and the
the frames shown are
noncoding) exon 1 of
referred to include all
alternative candidate

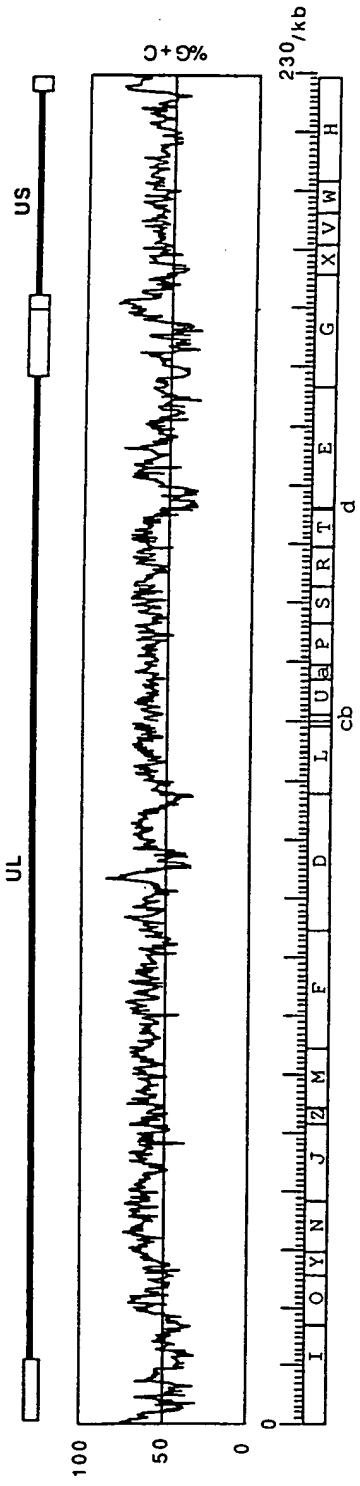
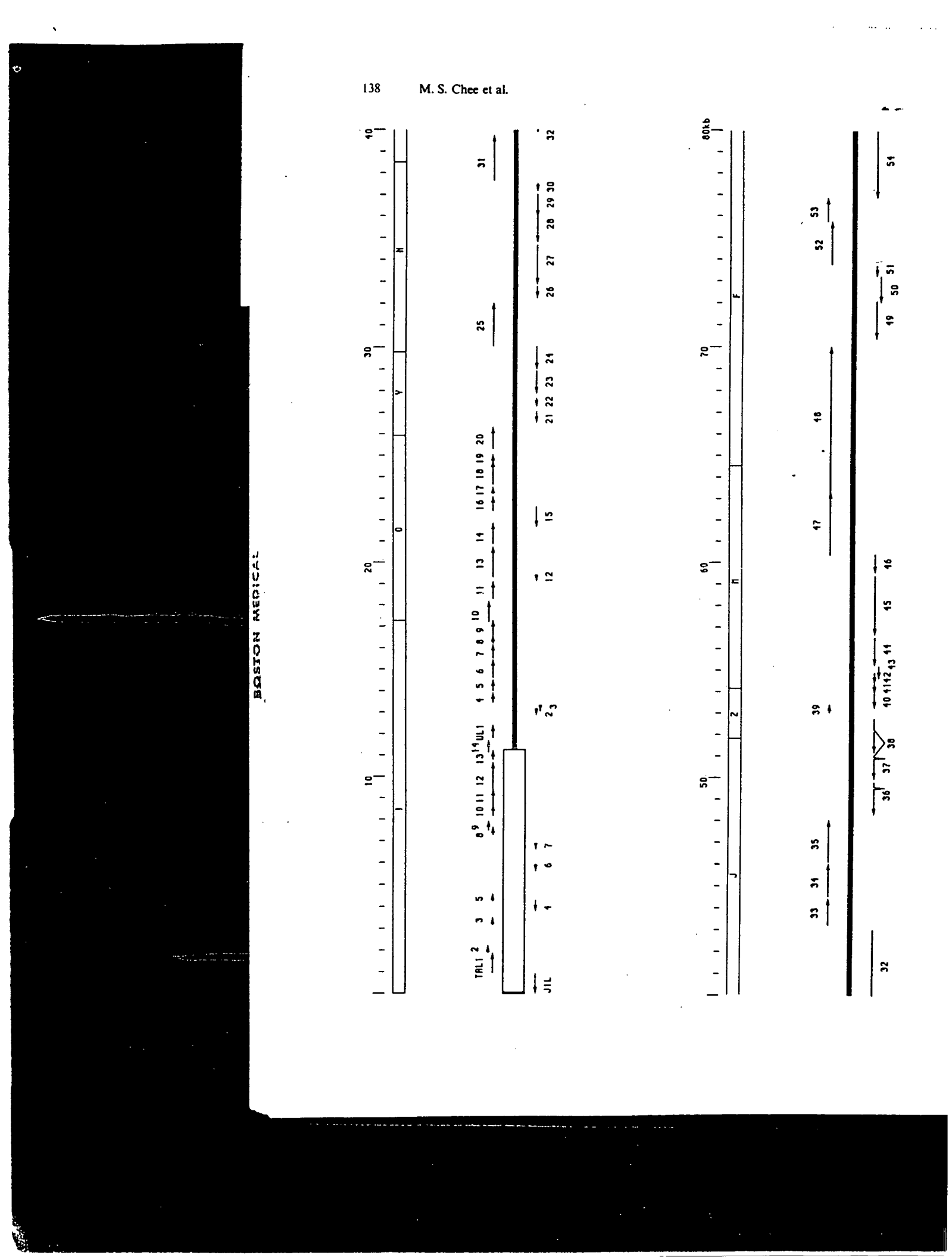


Fig. 2. Nucleotide composition of the HCMV strain AD169 genome. The % (G + C) content was plotted over the length of the genome using option 24 (plot base composition) of ANALYSEQ (STADEN 1986) with both span length and plot interval set at 201. The genome structure is shown above the plot, and a scale below. The orientation is that of the prototype isomer as indicated by the restriction map below the scale. The HCMV genome is relatively G + C rich (57.2% overall, 57.9% in UL, 55.7% in US, 49.9% in RL, 73.1% in RS). Within UL, marked variations in nucleotide composition are seen at either end in the *Hin*dIII fragments I, O, and E, and also in *Hin*dIIID. (see HONESS et al. 1989 for an analysis of dinucleotide frequencies)


10
20
30
40
50
60
70
80kb
I
O
Y
N
J
Z
F
TAL1
JIL
ULI

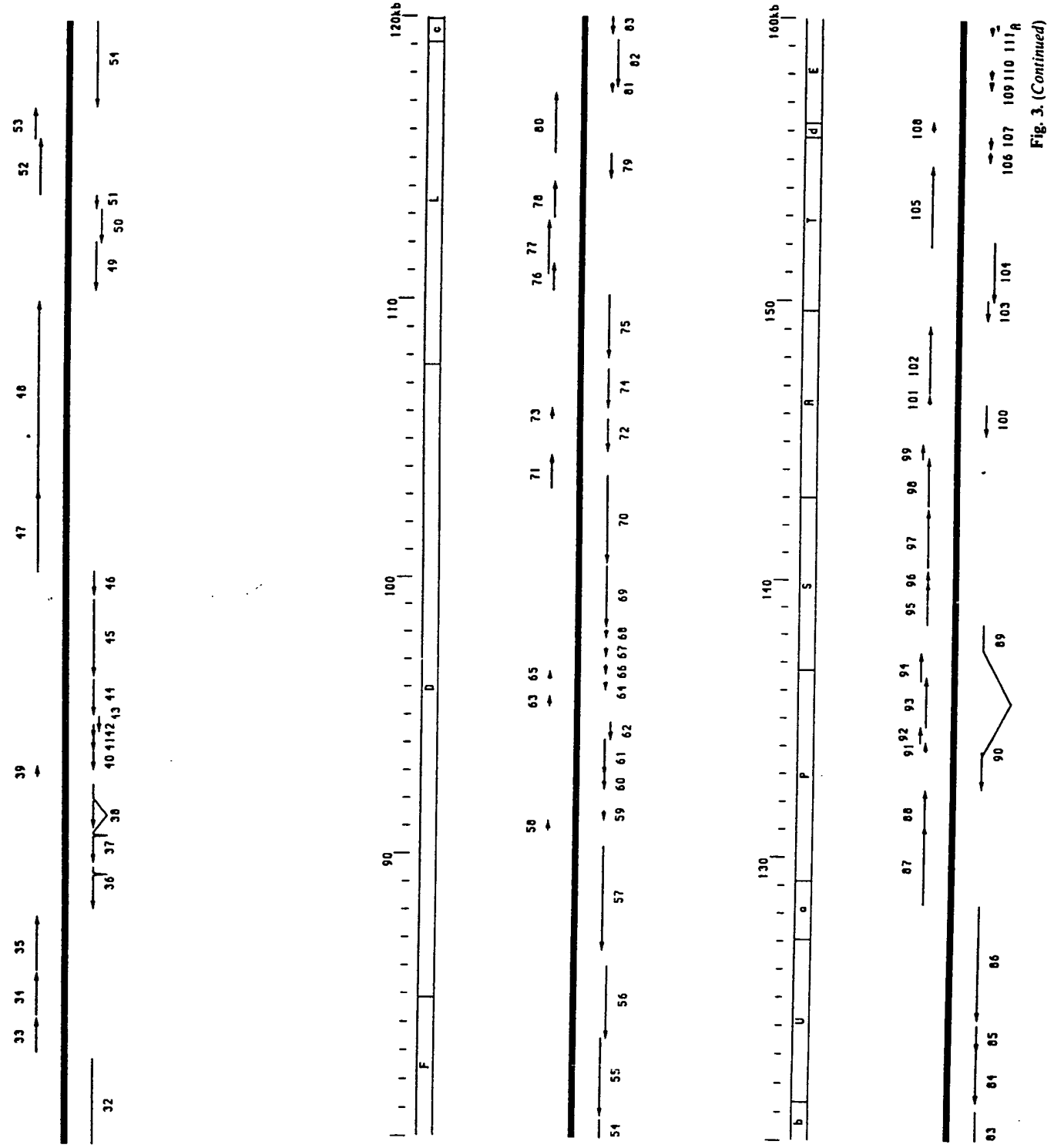
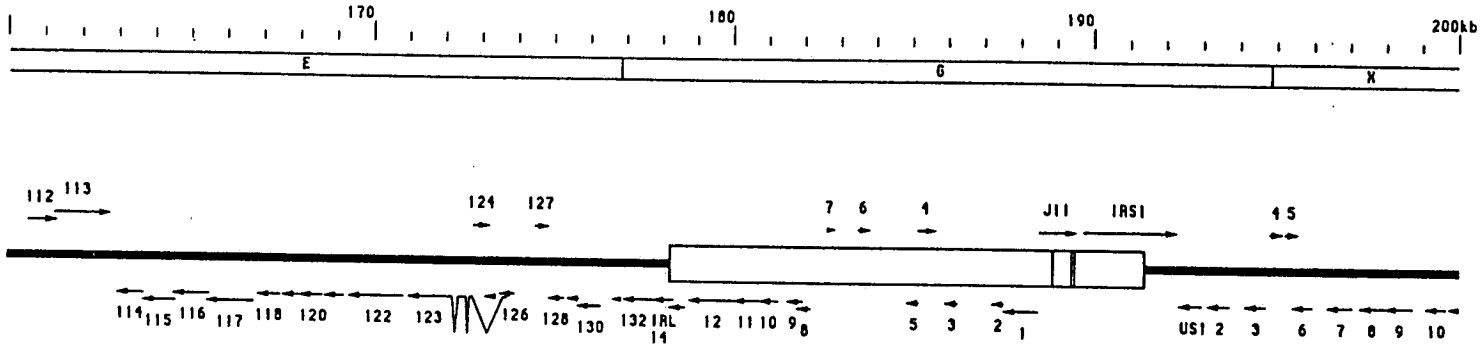
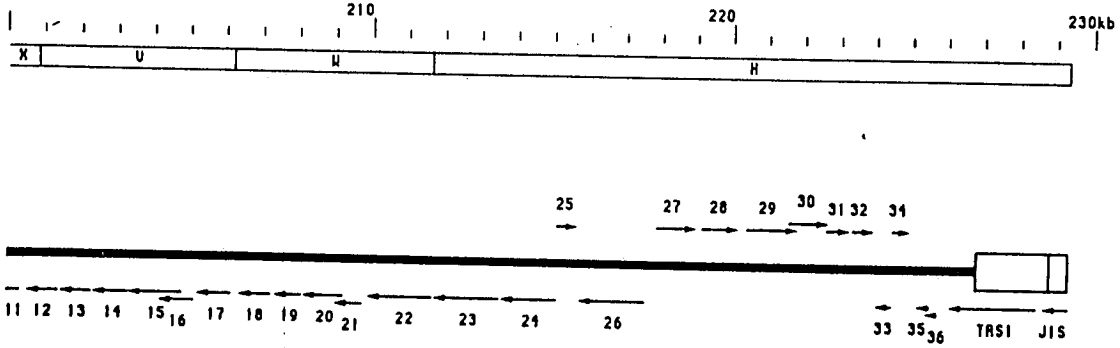


Fig. 3. (*Continued*)





4 Identification of Homologs

The HCMV protein sequences were screened against the PIR (release 19.0; GEORGE et al. 1986), and SWISSPORT (release 8.0; BAIROCH 1988) libraries using the FastA program of PEARSON and LIPMAN (1988). Searches were also performed against a herpesvirus protein library including HSV-1, VZV, and EBV sequences. In these library comparisons alignments were examined when optimized FastA scores of 90 or greater were obtained, although in some cases lower-scoring matches were also scrutinized. Some of the HCMV sequences match numerous reading frames as a result of compositional bias, which may be general throughout the sequence or localized. For example, glycine-rich stretches occur in a number of reading frames, including HCMV-UL44, 56, 102, 112, and TRS/IRS1. In most cases highly biased matches have been excluded. Sometimes, however, these similarities are likely to reflect functional similarities, if not homology. For example, HCMV-UL122, which encodes an immediate-early transactivator, is similar to HSV-IE110, also an immediate-early transactivator. The results of overall homology searches, motif searches (STADEN 1988), and comparisons of gene layout with EBV, VZV, and HSV-1 have been amalgamated in the compilation of human herpesvirus and cellular homologs. Functions ascribed to HCMV genes or their homologs are noted in Table 1. Homologies detected to the sequenced herpesviruses are shown in Table 2. A

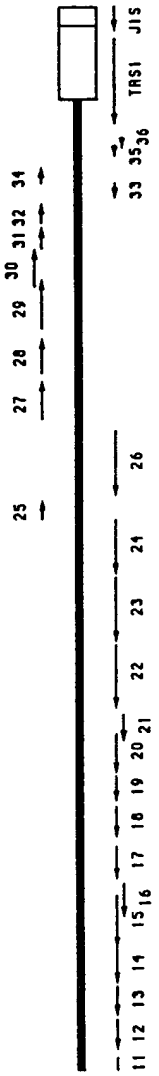


Fig. 3. A map of predicted open reading frames in HCMV strain AD169. Two hundred and eight individual frames are recognized, some of which are known to be spliced. The reading frame map is drawn in the prototype orientation below the *Hin*dIII restriction map. The diagram is scaled in kilobase pairs. Open reading frames which overlap on the same strand are displaced in the figure. Frames are numbered separately except for three genes for which splice sites have been precisely located (HCMV-UL36, UL37, and UL123) (KOUZARIDES et al. 1988; STENBERG et al. 1984, 1985), and one gene for which the splice sites are probably conserved with other herpesviruses (HCMV-UL89) (COSTA et al. 1985). Genes which may be spliced to upstream frames, but which are also capable of being initiated at a proximal ATG, are numbered separately (HCMV-UL36, UL38, UL122). Frames are designated TRL, IRL, UL, TRS, IRS, or US according to the region of the genome in which their 5′ ends are located, and each of these six sets is numbered from *1*. A frame which spans the DR1 repeats (Sect. 2) and hence is capable of crossing the genomic termini has been designated *J* (junction) *1*. Three manifestations of this frame which differ in their 5′ and 3′ termini occur, and are shown as *J1L*, *J1S*, and *J1I* (where *L*, *S*, and *I* denote long, short and internal respectively; see also Table 1). The "a" sequence is shown as a *thin vertical line* located within the repeats. The following frames have been included in place of longer overlapping frames; the names of the latter (not shown) are given *in brackets*, together with reasons for the substitution; the orientations of the substituted frames are indicated by the direction of numbering: *1*, J1L, and TRL1 (TRL1X, positions 291–1361; these frames occupy the region more completely, with minimal overlap. TRL1 has a proximal TATA box and a Kozak consensus ATG). [NB. J1L completely overlaps a frame equivalent to HKRFX (WESTON and BARRELL 1986) (not shown, positions 873–43)]; *2*, UL38 (UL38X, positions 51 098–52 141; third position G + C; see Sect. 5.3); *3*, UL106 (UL106X, positions 155 043–155 465; third position G + C); *4*, UL112 (UL112X, positions 161 638–160 466; third position G + C; mapping data; WRIGHT et al. 1988); *5*, UL123 (UL123X positions 172 331–172 816; overlaps major immediate-early gene exons 2 and 3); *6*, J1I and IRL1 (IRL1X, positions 189 176–188 106; see *1* above). US25X (former name HHRF1, positions 215 051–215 518; WESTON and BARRELL 1986) had an excessive overlap with US25 and was omitted without another frame being substituted in its place. The small frame UI111A (marked as *A*) was included because it has a Kozak consensus ATG, a transcript has been identified in the region, and it is a conserved feature of a transforming region in HCMVs Towne and AD169 (RAZZAQUE et al. 1988; JAHAN et al. 1989). The frame is one amino acid shorter than the Towne sequence, having a relative 3-bp deletion, but the predicted amino acid sequence is otherwise identical

Table 2. Homologs of HCMV-reading frames in the sequenced herpesviruses. Internal HCMV-related sequences as well as EBV, VZV, and HSV-1 homologs are listed, together with FastA scores (PEARSON and LIPMAN 1988). HCMV homologous families containing three or more sequences are indicated only in Table 1. We have found from experience that FastA scores above 100 are often significant, except when sequences are highly biased in composition. Homologs which were not identified by library searches, but which were inferred from their collinearity with other conserved frames, are scored as *P* (positionally conserved). Listings scored as *P*? should be regarded as tentative at best. Listings with a *question mark* and a FastA score show borderline similarity in the absence of supporting evidence and should be regarded as speculative. In most cases the highest scores above 90 were listed. Compositionally biassed matches were excluded for the following frames: HCMV-TRL/IRL4, TRL/IRL13, UL32, UL44, and UL113. Nomenclature for EBV, VZV and HSV-1 frames is conventional (BAER et al. 1984; DAVISON and SCOTT 1986; MCGEOCH et al. 1988a); the EBV sequence designated as LP (leader protein) is translated from the spliced EBNA2 mRNA (WANG et al. 1987)

		—Homologs						
Frame	HCMV	Score	EBV	Score	VZV	Score	HSV	Score
HCMVUL15			BCRF2?	93				
HCMVUL25	HCMVUL35	235					UL9?	87
HCMVUL35	HCMVUL25	235						
HCMVUL45			BORF2	151	VZV19	178	UL39	238
HCMVUL46			BORF1?	P	VZV20?	P	UL38?	P
HCMVUL47	HCMVUL86?	96	BOLF1?	P	VZV21?	P	UL37?	P
HCMVUL48			BPLF1	143	VZV22	P	UL36	144
HCMVUL49			BFRF2	249	VZV23	P	UL35	P
HCMVUL50			BFRF1?	P?	VZV24?	P	UL34?	P
HCMVUL51			BFRF1?	P?	VZV25	97	UL33	106
HCMVUL52			BFLF1	138	VZV26	179	UL32	207
HCMVUL53	HCMVUL69?	95	BFLF2	263	VZV27	99	UL31	141
HCMVUL54	HCMVUL130?	90	BALF5	343	VZV28	326	UL30	423
HCMVUL55			BALF4	720	VZV31	1061	UL27	1052
HCMVUL56	HCMVUL112?	95	BALF3	321	VZV30	290	UL28	323
HCMVUL57			BALF2	352	VZV29	220	UL29	298
HCMVUL61			LP?	181				
HCMVUL69	HCMVUL53?	95	BMLF1	P	VZV4	P	UL54	127
HCMVUL70			BSLF1	293	VZV6	302	UL52	405
HCMVUL71			BSRF1	92	VZV7?	P	UL51?	P
HCMVUL72			BLLF2	P	VZV8	P	UL50	88
HCMVUL73			BLRF1	134				
HCMVUL75	HCMVUL25?	90	BXLF2	217	VZV37	P	UL22	P
HCMVUL76			BXRF1	219	VZV35	151	UL24	132
HCMVUL77			BVRF1	316	VZV34	278	UL25	291
HCMVUL80			BVRF2	347	VZV33	177	UL26	243
HCMVUL82	HCMVUL83	325						
HCMVUL83	HCMVUL82	325						
HCMVUL85			BDLF1	P	VZV41	114	UL18	138
HCMVUL86	HCMVUL47?	96	BcLF1	1876	VZV40	767	UL19	1225
HCMVUL87			BcRF1	542	VZV38?	P	UL21?	P
HCMVUL89			BD/BGRF1	1181	VZV42/45	1104	UL15	1206
HCMVUL92			BDLF4	213				
HCMVUL93			BGLF1?	P	VZV43?	P	UL17?	P
HCMVUL94			BGLF2	241	VZV44	P	UL16	P
HCMVUL95			BGLF3	112	VZV46	P	UL14	P
HCMVUL97			BGLF4	157	VZV47	112	UL13	97
HCMVUL98			BGLF5	191	VZV48	78	UL12	140
HCMVUL99			BBLF1?	P	VZV49?	P	UL11?	P
HCMVUL100			BBRF3	417	VZV50	224	UL10	215
HCMVUL101			BBLF2?	P	VZV51?	P	UL9?	P
HCMVUL102			BBLF3?	P	VZV52?	P	UL8?	P
HCMVUL103			BBRF2	102	VZV53	91	UL7	121
HCMVUL104			BBRF1	357	VZV54	375	UL6	309
HCMVUL105			BBLF4	704	VZV55	642	UL5	598

Internal HCMV-related
h FastA scores (PEARSON
nces are indicated only in
significant, except when
d by library searches, but
scored as *P* (positionally
with a *question mark* and
nd should be regarded as
lly biassed matches were
32, UL44, and UL113.
984; DAVISON and SCOTT
in) is translated from the

Score	HSV	Score
	UL9?	87
178	UL39	238
P	UL38?	P
P	UL37?	P
P	UL36	144
P	UL35	P
P	UL34?	P
97	UL33	106
179	UL32	207
99	UL31	141
326	UL30	423
1061	UL27	1052
290	UL28	323
220	UL29	298
P	UL54	127
302	UL52	405
P	UL51?	P
P	UL50	88
P	UL22	P
151	UL24	132
278	UL25	291
177	UL26	243
114	UL18	138
767	UL19	1225
P	UL21?	P
1104	UL15	1206
P	UL17?	P
P	UL16	P
P	UL14	P
112	UL13	97
78	UL12	140
P	UL11?	P
224	UL10	215
P	UL9?	P
P	UL8?	P
91	UL7	121
375	UL6	309
642	UL5	598

Frame	HCMV	—Homologs Score	EBV	Score	VZV	Score	HSV	Score
HCMVUL112	HCMVUL56?	95						
HCMVUL114			BKRF3	545	VZV59	461	UL2	489
HCMVUL116			BDLF3?	128				
HCMVUL122							IE110?	90
HCMVUS2	HCMVUS3	169						
HCMVUS3	HCMVUS2	169						

survey of HCMV proteins including map assignments in the AD169, Towne, and Davis strain genomes has been conducted previously by LANDINI and MICHELSON (1988).

5 IE Genes

The activation of IE genes is the initial step in a viral program of gene expression. Northern hybridization studies have shown that transcription from the HCMV genome during the immediate early phase of productive infection is limited to several discrete loci, with the most active region located near one end of UL (DEMARCHI 1981; WATHEN and STINSKI 1982; MCDONOUGH and SPECTOR 1983; JAHN et al. 1984; WILKINSON et al. 1984). This major immediate-early (MIE) region has been studied in several CMV strains, and unlike the bulk of the CMV genome is CpG suppressed (HONESS et al. 1989). The MIE genes encode regulatory proteins, the expression of which requires only cellular factors, although virion components may also play a transactivating role (SPAETE and MOCARSKI 1985a; STINSKI and ROEHR 1985). More recently two other immediate-early loci have been sequenced and characterized in AD169 (KOUZARIDES et al. 1988; WESTON 1988).

5.1 MIE Gene Region

The first sequence data for this region were reported for HCMV Towne (STENBERG et al. 1984) and showed the four-exon arrangement of the major immediate-early (IE1) gene. Sequence analysis of the corresponding AD169 region revealed a similar arrangement with minor differences. Only two changes were observed at the amino acid level (AKRIGG et al. 1985). The organization of the equivalent murine CMV gene is grossly similar, but differs considerably at the sequence level (KEIL et al. 1987). Analysis of the HCMV IE promoter region exposed a complex array of 21-, 19-, 18-, and 16-bp repeats upstream of the TATA and CAAT boxes (THOMSEN et al. 1984; AKRIGG et al. 1985). The upstream sequence demonstrates a potent enhancer activity, detected by its ability to rescue enhancerless SV40 genomes (BOSHART et al. 1985). Homology with the core enhancer sequence TGGAAAG/TGGTTTG was

noted in the 18-bp repeats and potential Sp1-binding sites were also found. The enhancer binds cellular factors (GHAZAL et al. 1987, 1988) and dissection has shown that the 19-bp elements can mediate cAMP induction (FICKENSCHER et al. 1989; HUNNINGHAKE et al. 1989). Similar enhancers were also found in murine and simian CMVs (DORSCH-HASLER et al. 1985; JEANG et al. 1987). Nuclear factor 1 binding sites are associated with the enhancer region in both human and simian CMVs (HENNIGHAUSEN and FLECKENSTEIN 1986; JEANG et al. 1987).

STINSKI et al. (1983) recognized two further IE regions beginning immediately downstream of IE1. The IE2 region has more recently been called IE2a and a further region recognized as IE2b (HERMISTON et al. 1987; STENBERG et al. 1985). Under immediate-early conditions, transcription of the IE2a region starts mainly from the IE1 promoter and a set of alternatively spliced transcripts is produced. In the predominant species the IE2a exon (HCMV-UL122 in AD169) is fused to the first three exons of IE1. HCMV-UL122 encodes 494 amino acids following the splice acceptor. This is in agreement with the size predicted of the IE2a exon reported for the Towne strain by PIZZORNO et al. (1988). A 1.7-kb unspliced mRNA can also originate from a promoter proximal to the IE2a frame (which also contains a Kozak consensus ATG; KOZAK 1981). This transcript is more abundant at early and late times postinfection (STENBERG et al. 1985). The product of the IE2a frame may be involved in autoregulation (PIZZORNO et al. 1988). A minor transcript extending into the IE2b region has been diagrammed (HERMISTON et al. 1987). We are unable to correlate this with the AD169 sequence using the available information. However, a potential splice donor occurs before the UL122 termination codon, and a polyA signal at position 167 503 is consistent with the predicted end point of the Towne transcript. It is likely that the reading frames on either side of this signal, UL119 and UL118, are spliced together to encode a membrane glycoprotein.

5.2 HCMV US3 IE Gene

Sequencing of the US region of HCMV revealed an enhancer element containing five 18-bp repeats with homology to the MIE 18-bp repeats and the core enhancer element (WESTON 1988). These repeats were located in the region -80 to -270 of an RNA cap site in the HCMV-US3 (HQLF1) gene. In the region -340 to -600 a further set of six novel 11-bp repeats was found. A 275-bp fragment containing the 18-bp repeats enhanced expression in an orientation-independent manner in HeLa cells, with an efficacy equivalent to the SV40 enhancer (WESTON 1988), while the MIE enhancer 18-bp repeats have recently been shown to be involved in positive autoregulation by IE1 (CHERRINGTON and MOCARSKI 1989). The significance of the 11-bp repeats is unknown. However, a hexanucleotide consensus (TRTCGC) derived from these repeats was noted to occur in the MIE enhancer (WESTON 1988). Transcription from the HCMV-US3 reading frame associated with the enhancer is highly active at IE times and produces a set of differentially spliced transcripts. The protein-coding sequence of HCMV-US3 contains signal, anchor, and N-linked glycosylation sequences, is homologous to HCMV-US2 (HQLF2), and may also be related to the RLII and US6 gene families (Sect. 8).

5.3 UL37 IE Gene

A second UL IE transcription unit was identified in the region of the AD169 *Hin*dIII J and Z fragments (WILKINSON et al. 1984). The sequence of this region together with mapping data for three mRNAs has been published (KOUZARIDES et al. 1988). A 3.4-kb IE transcript was shown to be spliced from four exons and, like HCMV-US3, encodes a potential glycoprotein. This mRNA is 3′ coterminal with a 1.65-kb transcript which can be detected in the IE phase but is more abundant at the late stage of infection. The predicted product of the 1.65-kb mRNA is a member of the US22 homologous protein family (Sect. 7.2). A 1.7-kb transcript utilizing the same promoter as the 3.4-kb mRNA is most abundant at IE times but can also be detected late in infection. Of the mapped transcripts only this RNA contains the HCMV-UL38 (HZLF3) reading frame. However, expression of UL38 from this transcript would require the upstream UL37 exon 1 to be bypassed; alternatively, the frame may be read from an uncharacterized low-abundance transcript (KOUZARIDES et al. 1988). A 40-kDa protein synthesized in vitro from *Hin*dIII Z or J hybrid-selected mRNA is consistent with translation from UL38 (WILKINSON et al. 1984). Although a slightly longer reading frame completely overlaps UL38 on the opposite strand (UL38X, not shown), analysis of third position G + C contents suggests that of the two opposing frames UL38 is more likely to be coding (84.3% vs 62.8% G + C).

6 Early and Late Genes

Immediate-early proteins are required to activate genes which establish the early or delayed early (E or DE) phase of infection, the outcome of which is the replication of the viral genome. Late genes are expressed at high levels after DNA replication and are likely to encode most of the structural and assembly proteins of the virus. The distinction between E and late phases is blurred for some genes, and is further complicated by posttranscriptional regulation of gene expression (DEMARCHI 1983; GEBALLE et al. 1986a; GOINS and STINSKI 1986). In the following sections we attempt to correlate the available information on E and late genes with our sequence data. The organization of the following sections superficially resembles the viral timetable as convenient, but may be similarly inscrutable in places.

6.1 Major Early Transcripts

The most abundantly transcribed region of HCMV at early times postinfection is situated in the long repeats of the virus and encodes a 2.7-kb transcript of unknown function (GREENAWAY and WILKINSON 1987; HUTCHINSON et al. 1986; MCDONOUGH et al. 1985). An early transcript of similar size also originates in RL of HCMV Towne (WATHEN and STINSKI 1982), one copy of which can be deleted without compromising viability in cultured human fibroblasts (SPAETE and MOCARSKI 1987).

GREENAWAY and WILKINSON (1987) determined a 6220-bp sequence in HCMV AD169 which encompasses the gene for the 2.7-kb transcript. Their sequence is equivalent to positions 1635–7859 of Fig. 3 viewed in the opposite orientation. (We refer only to TRL sequence positions for clarity.) It contains two ambiguities and differs from our sequence at nine positions. However, only one of these is located within the major early transcription unit; the doublet CC beginning at position 3386 of GREENAWAY and WILKINSON (1987) is a triplet in our sequence. The open reading frame corresponding to the predicted translation product of the major 2.7-kb transcript as mapped by these authors is TRL/IRL4. The translational start is suggested to be the fourth ATG from the start of the transcript and occurs at position 4294 in our sequence. This is not a Kozak ATG in that it does not have a purine at -3 or a G at $+4$ (KOZAK 1981, 1982). However, two upstream ATG codons fit the Kozak consensus. The first has the sequence CGGATGG and is followed by a stop codon after seven amino acids. The second has the sequence GAGATGA and begins a 35-amino-acid reading frame. These codons have been shown to inhibit translation from a downstream AUG and may therefore be *cis*-regulatory signals (GEBALLE et al. 1986a; GEBALLE and MOCARSKI 1988). Upstream Kozak consensus ATGs precede a number of other HCMV genes, and suggest a general phenomenon in HCMV translational regulation. However, this role has yet to be demonstrated directly and so far no products have been found for the major early transcript. A less-abundant 2.0-kb transcript has been mapped immediately downstream of the 2.7-kb transcript in the Eisenhardt strain of HCMV (HUTCHINSON et al. 1986). The predicted polyadenylation site is conserved in AD169, beginning at position 6552 in our sequence. However, a similar-sized transcript was not detected (MCDONOUGH et al. 1985). It is also not possible to suggest a 5′ end from the Eisenhardt strain restriction map data. There are, however, no reading frames that might obviously be utilized in this region with the exception of TRL/IRL6. A minor 1.3-kb immediate-early RNA and a 1.2-kb late RNA have also been mapped to this general region (MCDONOUGH et al. 1985; HUTCHINSON et al. 1986); the latter is detected at early times postinfection but is most abundant in the late phase. The polyA signal for this message was located precisely in the Eisenhardt strain and begins at position 6365 of our sequence (HUTCHINSON et al. 1986). These authors also mapped the start of the transcript by nuclease protection and found no evidence for splicing. Further mapping and sequencing studies, the latter performed on genomic as well as cDNA clones, were used to predict a coding frame of 254 amino acids within the transcript (HUTCHINSON and TOCCI 1986). The region sequenced corresponds to positions 6300–7468 of Fig. 3 (displayed in the IRL orientation). However, in AD169 the 254-amino-acid reading frame is disrupted by three stop codons and two frameshifts relative to the Eisenhardt sequence and is identical in both repeats. Our data and those of GREENAWAY and WILKINSON are in agreement for the region spanned by the putative reading frame. We are unable to predict a reading frame which may be translated from this message in AD169. The first Kozak ATG occurs 164 nucleotides downstream of the transcription start predicted by HUTCHINSON and TOCCI (1986), but is followed by a stop codon after 42 intervening amino acid codons. Furthermore, although TRL/IRL7 is located in this message, it is over 500 bp from the predicted start. If

these differences between the Eisenhardt and AD169 strains are genuine, sequencing from other strains would be useful in assessing their biological relevance.

6.2 Enzymes of Nucleotide and DNA Metabolism

6.2.1 Nucleotide Metabolism

HONESS (1984) postulated that differences in overall base compositions between herpesvirus genomes reflect the ability of the viruses to modulate and utilize the nucleotide pool available for DNA synthesis. This hypothesis appears to be borne out in the case of the two closely related α-herpesviruses, HSV-1 and VZV. The latter is AT rich and encodes a thymidylate synthase, which does not have a homolog in the G + C rich HSV-1 genome (THOMPSON et al. 1987; MCGEOCH et al. 1988a). A parallel exists in the less closely related γ-herpesviruses Epstein-Barr virus (EBV) and herpesvirus saimiri (HVS); the latter A + T rich virus encodes thymidylate synthase and dihydrofolate reductase, which both seem to be absent from the G + C rich EBV (HONESS et al. 1986; TRIMBLE et al. 1988; BAER et al. 1984). All four viruses also encode deoxyribonucleoside kinases, and hence can utilize the salvage pathway of dNTP synthesis (MCKNIGHT 1980; DAVISON and SCOTT 1986; LITTLER et al. 1986; GOMPELS et al. 1988a). These enzymes differ in their substrate specificity and their main role might be to allow the exploitation of specific cell types, such as may occur in latency. Genes for ribonucleotide reductase, a key enzyme in deoxyribonucleotide synthesis, have been found in HSV, VZV, and EBV as well as other herpesviruses, but have not so far been identified in HVS (GIBSON et al. 1984; DAVISON and SCOTT 1986; NIKAS et al. 1986). The HCMV genome is relatively G + C rich (Fig. 2) and it will be of interest to determine if its complement of enzymes is consistent with the theory of HONESS (1984). HCMV does not appear to encode a thymidine (deoxyribonucleoside) kinase (TK); the position in the AD169 genome equivalent to the TK locus in other herpesviruses is deleted relative to the other herpesviruses (Fig. 3). However, HCMV is sensitive to the nucleoside analog DHPG, and a resistant mutant of AD169 has been isolated which accumulates less of the triphosphate form of the drug (BIRON et al. 1986). This may indicate that a deoxyribonucleoside kinase is encoded at some other locus.

The partial conservation of a ribonucleotide reductase (RR) homolog is more puzzling. Mammalian cells contain an iron-tyrosyl radical enzyme, which is the type found in herpesviruses (SJOBERG et al. 1985; REICHARD 1989). The enzyme has an $\alpha_2\beta_2$-structure; the HCMV-UL45 gene product is homologous to the α-(large) RR subunit, and HCMV-UL45 is positionally conserved with the gene for this subunit in other herpesviruses. However, the gene for the β-(small) subunit does not appear to be conserved; HCMV-UL44 is positionally analogous to the small RR gene in other herpesviruses but encodes a set of late DNA-binding proteins (see Sect. 6.5). The small subunit contains the active tyrosyl radical and would be essential for function. Thus it is not clear at present if HCMV is capable of expressing a fully active ribonucleotide reductase. Although we have used loosely defined motifs to search all the predicted reading frames for a potential active site, no obvious

candidates were identified. Several explanations could account for this. For example, if HCMV-UL45 is functionally conserved with the large subunit, it might usurp the place of its cellular counterpart which mediates allosteric control as well as being involved in catalysis. Herpesviral reductases appear to be unregulated, indicating that the function is either unnecessary or perhaps detrimental in the viral context (LANIKEN et al. 1982; AVERTT et al. 1983). It is also possible that synthesis of one or both of the cellular subunits is upregulated during viral infection (STINSKI 1977). The genes for the human RR subunits are unlinked; the α-subunit gene is on chromosome 11 (ENGSTROM et al. 1985), and the β-gene on chromosome 2 (YANG-FENG et al. 1987). Finally, it is worth mentioning that another key allosteric enzyme of nucleotide metabolism is dCMP deaminase; this enzyme converts dCMP to dUMP, which is the substrate for thymidylate synthase. Hence it might be an appropriate enzyme for herpesviral repertoires, particularly those which have devolved to an A + T bias.

6.2.2 *DNA Replication*

A set of seven HSV-1 genes has been shown to be essential for the replication of an HSV-origin-containing plasmid (WU et al. 1988; MCGEOCH et al. 1988b). The HCMV homologs of four of these have been identified by sequence analysis. HCMV-UL54 encodes the DNA polymerase (KOUZARIDES et al. 1987a; HEILBRONN et al. 1987) and HCMV-UL57 the major DNA-binding protein (MDBP). The latter = ICP8
sequence shows 72% identity over a length of 1160 aligned amino acids to the MDBP of simian CMV (Colburn) (ANDERS and GIBSON 1988; ANDERS and GIBSON, personal communication). HCMV-UL105 encodes a homolog to HSV-UL5, which is probably a helicase enzyme (CRUTE et al. 1988, 1989). Helicases belong to a superfamily of proteins with functions in replication and/or recombination (HODGMAN 1988). A nucleotide-binding site in UL105 (MARTIGNETTI 1987), of the type GxxGxGK (where x = any amino acid), is common to the other members of the superfamily. HCMV-UL70 is the fourth HCMV gene with an obvious replication gene counterpart, in HSV-UL52. The product of HSV-UL52 is part of a helicase-primase complex in HSV-1-infected cells which also contains the HSV-UL5 and UL8 proteins (CRUTE et al. 1989). HCMV genes UL102 and UL101 are positionally equivalent to HSV-UL8 and UL9 respectively, although they show no clear-cut homology. However, HCMV-UL102 is a similar length to HSV-UL8 (798 and 750 residues respectively). HSV-UL9 encodes an origin-binding protein (OLIVO et al. 1988), and the positive identification of its HCMV counterpart may require the identification of an HCMV origin of replication.

6.2.3 *DNA Repair*

The gene for uracil-DNA glycosylase, which is involved in base excision repair, was identified in HSV-2 and is conserved in the sequenced herpesviruses (WORRAD and CARADONNA 1988; BAER et al. 1984; DAVISON and SCOTT 1986; MULLANEY et al. 1989). The corresponding HCMV-reading frame is HCMV-UL114, which is the last frame at this end of UL with detectable homology to sequenced human herpes-

viruses. A dUTPase gene is also conserved in herpesviruses, albeit less well than uracil-DNA glycosylase (PRESTON and FISHER 1984; DAVISON and SCOTT 1986; BAER et al. 1984). The HCMV homolog is HCMV-UL72.

6.2.4 Deoxyribonuclease

A deoxyribonuclease gene found in HCMV appears to be ubiquitous in herpesviruses, as homologs are found in HHV-6 (LAWRENCE et al., unpublished results), EBV (ZHANG et al. 1987), HSV (MCGEOCH et al. 1986), and VZV (DAVISON and SCOTT 1986). The role of this enzyme is currently unknown, but it may be involved in cleavage of viral concatemers and/or the processing of genome termini (CHOU and ROIZMAN 1989).

6.3 Phosphotransferase

The putative phosphotransferase encoded by HCMV-UL97 is conserved in the human herpesviruses and distantly related to the protein kinase family (CHEE et al. 1989a; SMITH and SMITH 1989). Interestingly, some of the most conserved amino acids in protein kinases are variant in the herpesvirus sequences. One motif where these differences occur is shared with bacterial phosphotransferases, which vary at the same amino acid positions as do the herpesvirus proteins (BRENNER 1987). Hence it remains to be shown if HCMV-UL97 and its homologs are in fact conventional kinases. Whatever its specific role, the preservation of this gene in all of the recognized herpesvirus lineages and HHV-6 implies an important or indispensable contribution to the viral life cycle. None of the other HCMV-reading frames we have screened have detectable homology to known protein kinase motifs, which are seen in the α-herpesvirus US-encoded kinases (MCGEOCH and DAVISON 1986).

6.4 Early Phosphoprotein Genes

The gene for a set of phosphoproteins sharing a common N-terminus has been mapped by WRIGHT et al. (1988). These authors mapped the termini of two spliced 2.2-kb early transcripts, raised an antiserum against a synthetic peptide predicted from a 5'-terminal portion of the 5'-exon sequence (KOUZARIDES et al. 1983; RASMUSSEN et al. 1985a) and used this to detect four proteins of 34, 43, 50, and 84 kDa in infected cells (WRIGHT et al. 1988). Pulse-chase experiments did not suggest that any of the proteins were derivative in nature. Although the mapping data are as yet incomplete, it would thus appear that all four proteins are coded in alternatively spliced mRNAs sharing a 5' exon. This exon corresponds to UL112 in our sequence. A 279-bp portion of the UL113 frame (positions 161 503–161 781) is flanked by potential acceptor and donor sites, and may correspond to a 280-bp exon mapped by STAPRANS and SPECTOR (1986). The downstream exons may also be derived from UL113, which extends to position 162 797. A polyA signal begins at position 162 909, but there is an alternative polyA sequence coinciding with the end

of UL113 (ATTAAA, beginning at position 162 796). It therefore seems likely that one or both of these signals indicates the end of the transcription unit. The four proteins were found to be predominantly contained in the nuclear fraction of infected cells, and were not shown to be virion structural proteins in preliminary studies (WRIGHT et al. 1988).

6.5 Late DNA-Binding Proteins

Mocarski and coworkers utilized immunological screening of a λgt11 expression library to map a group of proteins known as the ICP36 family to the HCMV-UL44-reading frame (MOCARSKI et al. 1985; LEACH and MOCARSKI 1989). The ICP36 proteins gravitate to the nucleus, include phosphorylated and glycosylated species, and are DNA-binding proteins (PEREIRA et al. 1982; GIBSON 1983; MOCARSKI et al. 1985). Regulation of HCMV-UL44 gene expression is manifested in both early and late transcription from different TATA boxes, and delayed translation of early message (LEACH and MOCARSKI 1989; GEBALLE et al. 1986b). The significance of this complex control is unclear, although it is interesting that the 3'-end of the reading frame is overlapped by a gene encoding a small RNA in the same orientation. This gene is probably transcribed by RNA polymerase III (MARSCHALEK et al. 1989).

6.6 Capsid Proteins

The gene for the major capsid protein (MCP) was identified by sequence homology to the MCP sequences of other human herpesviruses and the assignment confirmed immunologically (CHEE et al. 1989b). The MCP is encoded by the HCMV-UL86 reading frame. Homology searches show that the predicted protein sequence of another frame, HCMV-UL47, is similar to a region of the human herpesvirus major capsids corresponding approximately to positions 1080–1170 of Fig. 3 in (CHEE et al. 1989b). Although this match may be fortuitous, the alignment of HCMV-UL47 to conserved capsid sequences makes it of interest. However, the sequence is not obviously conserved in the EBV, VZV, and HSV-1 reading frames collinear with HCMV-UL47.

A second capsid protein, which is a constituent of incomplete capsids, has been mapped in the UL region of three CMV strains (ROBSON and GIBSON 1989). Several lines of evidence implicate this protein in DNA packaging and/or capsid assembly (PRESTON et al. 1983; IRMIERE and GIBSON 1985; LEE et al. 1988; RIXON et al. 1988). The gene for the putative assembly protein is conserved in the human herpesviruses, and is predicted to encode proteins of 635, 605, 605, and 708 amino acids in HSV, VZV, EBV, and HCMV respectively (MCGEOCH et al. 1988a; DAVISON and SCOTT 1986; BAER et al. 1984) (Table 1). The sequence of a 1-kb cDNA derived from the Colburn strain of CMV shows homology only to the 3' half of HCMV-UL80, consistent with the 37-kDa size of the Colburn strain assembly protein which is probably processed at the carboxy terminus (ROBSON and GIBSON 1989). A larger transcript of 1.8-kb is also encoded at this locus. The 5' portion of the HCMV-UL80

frame is conserved in the other sequenced human herpesviruses. It thus seems likely that at least two seperate proteins are encoded by HCMV-UL80, with a TATA box at position 115992 being used to produce the assembly protein transcript (ROBSON and GIBSON 1989). This TATA box is located within 15 bp which are identical in Colburn and AD169 (NECKER et al. 1988 cited in ROBSON and GIBSON 1989). It is also noteworthy that the ATG downstream of this TATA box does not fit the Kozak consensus in either of the two CMV sequences. In contrast to the major DNA-binding protein (Sect. 6.2.2), the sequences for the putative assembly protein are quite divergent. The Colburn sequence from the first methionine of the predicted cDNA reading frame exhibits approximately 40% identity to the carboxy-terminal 371 amino acids of HCMV-UL80.

6.7 Structural Phosphoprotein Genes

HCMV virions contain three main phosphoproteins which appear to be located in the virion tegument (ROBY and GIBSON 1986). The largest of these is approximately 150 kDa in size, constitutes approximately 20% of virion protein content (IRMIERE and GIBSON 1983), and is also modified by O-linked glycosylation (BENKO et al. 1988). A 6360-bp region containing the pp150 gene sequence (which corresponds to the reading frame HCMV-UL32) has been published and spans positions 37157–43516 of Fig. 3 viewed in the opposite orientation. A late 6.2-kb mRNA was mapped in this region, and its termini delineated. Some processing at an alternative polyA site (ATTAAA) downstream of the orthodox signal was demonstrated. The major RNA species is predicted to encode pp150 although a range of smaller RNA species was also detected (JAHN et al. 1987).

The two other major phosphoproteins located in virions are pp71 and pp65, also known as the upper and lower matrix phosphoproteins respectively. The 65-kDa phosphoprotein is also glycosylated (CLARK et al. 1984; PANDE et al. 1984), and pp71 may be similarly modified. The genes for pp65 and pp71 are located in the *Hin*dIII L, c, b region of the genome and correspond to reading frames HCMV-UL83 and UL82 respectively. The sequence of a *Hin*dIII/*Bgl*II fragment containing these genes has been reported, and corresponds to nucleotides 117276–121377 of Fig. 3 viewed in the opposite orientation (RUGER et al. 1987). The published sequence is in error; position 212 (121166 in the genome) is shown as a G but should be read as a C. This change does not affect the predicted coding sequences. Two transcripts which appear to be 3′ coterminal were mapped in this region. They are an abundant 4-kb mRNA and a low-level 1.9-kb mRNA. The 5′ ends of both transcripts have been located, but surprisingly no TATA box is proximal to the 4-kb transcription unit (RUGER et al. 1987). The 4-kb message should encode pp65, while the shorter mRNA would allow pp71 to be translated. The mRNA encoding pp65 (ICP27) in HCMV Towne appears to be produced efficiently both early and late in infection, but is not translated at high levels until the late phase (GEBALLE et al. 1986b; but see DEPTO and STENBERG 1989). The gene sequences for two further structural phosphoproteins have been reported (MEYER et al. 1988; DAVIS and HUANG 1985). The data of MEYER et al. (1988) represent positions 143791–145191 of our sequence in the *Hin*dIII R

Table 3. HCMV glycoprotein genes. A compilation of signal and anchor sequences and numbers of possible N-linked glycosylation sites in 54 reading frames. The selection of frames was based on criteria defined by McGeoch (1985). A *questionmark* after the number of NXT/S sites indicates that at least one of these sites is located on the putative cytoplasmic face of the sequence. Twenty-two of the frames lack at least a signal or an anchor sequence. Many of these may represent glycoprotein exons (Table 1), while some may encode unusual or non-N-linked glycoproteins like the pseudorabies gp50 (Petrovskis et al. 1986). It is also possible that some of the potential glycoproteins may be fixed to membranes by glycosyl-phosphatidylinositol anchors (Ferguson and Williams 1988)

Frame	Strand	Signal	NXT/S	Anchor
HCMVTRL/IRL3		MYCFLFLQKDTFFHEQFLARRRHAE		IGVLVVVCGFYFFLYLSMTVFLFFVLIII
HCMVTRL/IRL10		MYPRVMHAVCFLALSLVSYVAVCAE	4	EPITMLGAYSAWGAGSFVATLIVLLVVFFVIYAR
HCMVTRL/IRL11		MQTYSTPLTLVIVTSLFLFTTQGSS	3	HCAWVSGMMIFVGALVICFLR
HCMVTRL/IRL12		MRVACRRPHHLTYRHTAYTIIIFYI	23	SRTVWTIVLVCMACIVLFFAR
HCMVTRL/IRL13		MDWRFTVMWTILISALSESCNQTCS	9	
HCMVTRL14			3	HAVWAGVVVSVALIALYMGSH
HCMVUL1		MGMQCNTKLLLPVALIPVVIILIGT	9	HAGWAAAVVTVIMIYVLIHFNVPATLR
HCMVUL2	C		-	RGIFLITLVIWTVVWLKLLR
HCMVUL4		MVMMLRTWRLLPMVLLAAYCYCVFG	9	
HCMVUL5			-	HTTWVTGFVLLGLLTLFASLFR
HCMVUL6		MHAKMNGWAGVRLVTHCLNTRSRTY	10	LAFTYGSWGVAMLLFAAVMVLVD
HCMVUL7			11	HLALVGVIVFIALIVVCIMGWWK
HCMVUL8			3	HYSWMLIIAIILIIFIIICLR
HCMVUL9		MYRYTWLLWWITILLRIQQFFYQWWK	5	HTMWIIPLVIVTTIIVLICFK
HCMVUL10			3	HSAWILIVIIIIIVVILFFFK
HCMVUL11		MLLRYITFHREKVLYLAIACFFGIY	5	HALWVLAVVIVIIIIIIFYFR
HCMVUL12	C	MLLVFLGPVNSFMKGIRDVGFGKPP	1	
HCMVUL13		MLWAHCGRFLRYHLLPLLLCRLPFL	2	
HCMVUL14		MWSRVVFLRSSETQTGMGGGRLPPL	2	KIGLLAAGSVALTSLCHLLCYWCSE
HCMVUL16		MERRRGTVPLGWVFFVLCLSASSSC	8	DIVLVSAITLFFFLLALR
HCMVUL18		MMTMWCLTLFVLWMLRVVGMHVLRY	13	RYNTMTISSVLLALLLCALLFAFLH
HCMVUL20		MLGIRAMLVMLDYYWIQLITNNDTR	13	RYMYLFSVSCAGITGTVSIILVSLSLLILICYYR
HCVMUL37EX3	C		16	HWALLSICTVAAGSIALLSLFCILLIGLR
HCMVUL37EX1	C	MSPVYNLLGSVGLLAFWYFSYRWI	-	
HCMVUL40	C	MNKFSNTRIGFTCAVMAPRTLILTV	3	ETWAMYTVGILALGSFSSFYSQIAR
HCMVUL42	C		1?	KWTFALLVVAILGIIFLAVVFTVVINR
HCMVUL50	C		2	RFATLGPLVLALLLVLALLWR
HCMVUL55	C	MESRIWCLVVCVNLCIVCLGAAVSS	21	KNPFGAFTIILVAIAVVIITYLIYTR
HCMVUL67	C	MVRSLEEIIYIIYSDDSVVNISLAS	3	
HCMVUL73		MEWNTLVLGLLVLSVVAESSGNNSS	3	ELSLSSFAAWWTMLNALILMGAFCIVLR
HCMVUL74	C	MGRKEMMVRDVPKMVFLISISFLLV	20	

HCMVUL75	C	MRPGLPPYLTVFTVYLLSHLPSQRY	5	RLLMMSVYALSAIIGIYLLYR
HCMVUL118	C		8	RLLAYGVLAFLVFMVIILLYVTYMLAR
HCMVUL119	C	MCSVLAIALVVALLGDMHGVKSST	4	
HCMVUL120	C	MYRAGVTLLVVAVVSLGRWDVVTMA	9?	RAFMIVILTQVVFVVFIINASFIWSWTFR
HCMVUL121	C	MWGCGWSRIIVLLPLMCMALMARGT	1	DLGLLYAVCLILSFSIVVAALWK
HCMVUL124		MERNSLLVCQLLCLVARAAATSTAQ	3	DTYPTATALCGTLVVVGIVLCLSLASTVR
HCMVUL129	C		-	RIFMIVCLWCVWICLSTFLIAMFH
HCMVUL130	C	MLRLLLRHHFHCLLLCAVWATPCLA	3	
HCMVUL132	C	MPAPRGLLRATFLVLVAFGLLLHID	3?	EIMKVLAILFYIVTGTSIFSFIAVLIAVVYSSCCK
HCMVUS2	C	MNNLWKAWVGLWTSMGPLIRLPDGI	1	HVAWTIVFYSINTLLVLFIVYVTVD
HCMVUS3	C	MKPVLVLAILAVLFLRLADSVPRPL	1	RTLLVYLFSLVVLVLLTVGVSAR
HCMVUS6	C	MDLLIRLGFLLMCALPTPGERSSRD	1	HGFFAVTLYLCCGITLLVVILALLCSITYE
HCMVUS7	C	MRIQLLLVATLVASIVATRVEDMAT	2?	RWLTIILYVFMWTYLVTLLQYCIVR
HCMVUS8	C	MRRWLLVGLGCCWVTLAHAGNPY	2	LELGVVIAICMAMVLLLGYVLAR
HCMVUS9	C	MILWSPSTCSFFWHWCLIAVSVLSS	2	HVALFSFGVQVACCVYLR
HCMVUS10	C	MLRRGSLRNPLAICLLWWLGVVAAA	2	DYGAILKIYFGLFCGACVITR
HCMVUS11	C	MNLVMLILALWAPVAGSMPELSLTL	1	KSAQYTLMMVAVIQVFWGLYVK
HCMVUS34		MNLEQLINVLGLLWIAARAVSRVG	4	

fragment and show the gene for a 28-kDa protein encoded by a late 1.3-kb RNA. MARTINEZ et al. (1989) and MARTINEZ and ST. JEOR (1986) mapped a 25-kDa protein to the same locus and assigned a 1.6-kDa late mRNA as the message. These RNAs are likely to be initiated from one or both of two TATA boxes proximal to HCMV-UL99. An HCMV Towne 1.4-kb late mRNA localized to this region may also denote HCMV-UL99 (PANDE et al. 1988). However, the Towne protein migrates as a 32-kDa protein. If the same frame is in fact being used, nontrivial explanations for the difference could be invoked at the genetic, transcriptional, and protein-processing levels. It is interesting to note that a minor 27-kDa species was detected by PANDE et al. (1988) in infected cells and virions.

An example of a phosphoprotein gene that appears not to be conserved between HCMVs Towne and AD169 was mapped and sequenced from passage 36 of HCMV Towne (DAVIS et al. 1984; DAVIS and HUANG 1985). This gene encodes an abundant late transcript, and immunological evidence suggests that its product is a 67-kDa nonglycosylated phosphoprotein found in virions. The sequenced fragment corresponds very approximately to a region of AD169 *Hin*dIII D beginning at about position 95 500. There appear to be significant differences between the two genomes in this region. These include numerous point and frameshift mutations and a deletion of 61 bp in Towne relative to AD169. A consequence of some of these differences is the disruption of the putative Towne reading frame in AD169, although a portion of the predicted phosphoprotein sequence is preserved in HCMV-UL65. The reported sequence was not determined fully on both strands, and not all sequenced fragments were shown to be contiguous. Hence further comparative sequence analysis and transcript mapping will be necessary before these findings can be interpreted unambiguously, particularly as the equivalent region in AD169 contains some potential splice sites. A gene which is posttranscriptionally regulated by an mRNA 3'-end processing event was partially sequenced and shown to contain a potential stem-loop structure (GOINS and STINSKI 1986). This sequence maps to positions 96 753–97 076, and may therefore correspond to the 3' end of a transcription unit spanning HCMV-UL65. The putative stem-loop structure in the Towne sequence is conserved in AD169, although there are three deletions relative to AD169 clustering in the 3'-terminal 25 nucleotides of the published sequence.

6.8 Surface Glycoproteins

The importance of glycoproteins as surface antigens has made the major HCMV glycoproteins a focus for characterization and functional studies. A total of 54 reading frames have now been found in the sequence that have charcteristics of glycoprotein genes or of exons of glycoprotein genes. These are presented in Table 3, which shows the predicted signal sequences, the number of N-linked glycosylation sites, and the anchor sequences. Twenty-two of these frames lack either a signal or an anchor. In the following sections we consider two immunologically important glycoproteins, and two which have homology to host immunoglobulin superfamily proteins. Known IE glycoprotein genes and glycoprotein gene families are considered separately in Sects. 5 and 7 respectively.

6.8.1 Glycoproteins B and H

There are seven virion glycoproteins encoded by HSV-1 and one putative glycoprotein (US5) predicted from the sequence (McGeoch et al. 1988a). Of these five have counterparts in the sequence of VZV (Davison and Scott 1986) and only two in the genome of EBV (Baer et al. 1984). In addition, EBV has the gp350/220 (BLLF1a, b), BILF1, and BLRF1 glycoproteins. The latter has a homolog in HCMV-UL73. Of the other herpesvirus glycoproteins, only homologs to gB (HCMV-UL55) (Cranage et al. 1986; Kouzarides et al. 1987b; Mach et al. 1986) and gH (HCMV-UL75) (Cranage et al. 1988; Pachl et al. 1989) have been found in the HCMV sequence, and so gB and gH are common to all of the well-studied herpesviruses. The conservation of gH in distantly related herpesviruses (Gompels et al. 1988b) and the production by an HSV-1 ts mutant of noninfectious virus lacking gH (Desai et al. 1988) underpin the substantial body of immunological evidence that gH is essential for virus infectivity. Monoclonal antibodies to HCMV gH can neutralize virus in vitro unassisted by complement (Rasmussen et al. 1984; Cranage et al. 1988). Antibodies to gB are also able to neutralize virus in vitro, but require complement (Cranage et al. 1986). A virion envelope glycoprotein complex has been shown to contain gB, but the structural nature of this entity awaits definition (see, for example, Farrar and Greenaway 1986; Gretch et al. 1988a). The unmodified gB precursor in AD169 is predicted to be 102 kDa in size. This is processed and glycosylated to give a 145-kDa species which is proteolytically cleaved to produce a 55-kDa species, both of which can be detected in infected cells. However, the residual 90-kDa amino-terminal cleavage product is not detected (Cranage et al. 1986). The site of cleavage has been mapped to Arg_{450} in the gB of HCMV Towne and by analogy processing of the AD169 gB is likely to occur after Arg_{459} (Spaete et al. 1988). These authors also compare the gene and protein sequences of gB and find identities of 94% and 95% respectively between the two HCMV strains. (A similar level of conservation is found between the gH sequences of these strains; Pachl et al. 1989.) There appear to be noteworthy differences in the kinetics of gB transcription in these two strains. The AD169 gB transcripts are produced late in infection (Kouzarides et al. 1987b) while the Towne gB mRNA is of the early class. However, in HCMV Towne infected cells gB is not detected immunologically until late in infection (Rasmussen et al. 1985b), implying that the two strains might use different strategies to achieve a similar result in the regulation of gB expression.

6.8.2 HLA Homolog

The identification of an HCMV glycoprotein with homology to class I major histocompatibility (MHC) antigens has implications for host-virus interactions (HCMV-UL18, Beck and Barrell 1988). The crystal structure of a human class I histocompatibility molecule (HLA-A2) has been solved (Bjorkman et al. 1987a), making it possible to predict that the HLA homolog is likely to have three extracellular domains analogous to the class I α1-, α2-, and α3-domains. The latter contains a β_2-microglobulin (β_2m)-binding loop which is partially conserved in the

HCMV sequence (BECK and BARRELL 1988). In cellular HLA molecules, the α3-domain and associated β_2m are both β-sandwich structures surmounted by the α1- and α2-domains which each contain a long α-helical region. A groove between these helices forms an antigen-binding cleft while surface residues may be involved in binding to a T-cell receptor (TCR) (BJORKMAN et al. 1987b). In contrast to the cellular sequences, both the α1- and α2-domains in the HCMV homolog are potentially heavily glycosylated as they contain a total of ten NXS/T motifs. Three or four of these motifs are located in the predicted helical and interhelical domains and hence might have a direct bearing on any antigen or TCR binding ability of the molecule. The protein expressed in vaccinia recombinants is in fact heavily glycosylated (H. BROWNE and A. MINSON, personal communication). In light of recent evidence that murine CMV can prevent the association of specific viral antigens with MHC (DEL VAL et al. 1989), a role for the HCMV HLA homolog in infected cells can be proposed. That is, the viral protein may compete with cellular HLA for the binding of one or more specific viral antigens, and consequently interfere with their presentation on the cell surface (TOWNSEND et al. 1989). While it is also possible that β_2m binding in the HCMV tegument may be due to the HLA homolog, no evidence for a link between the two has yet been presented (STANNARD 1989; GRUNDY et al. 1987a, b). Whatever the function of the protein, when co-expressed with β_2m from vaccinia vectors it is capable of associating with β_2m, which can then be detected on the cell surface (H. BROWNE and T. MINSON, personal communication). Finally, it should be noted that this gene does not have a homolog in the other sequenced human herpesviruses, and is found in a region which appears to be unique to β-herpesviruses.

6.8.3 T-Cell Receptor Homology

Even more provocative than the identification of a HLA homolog is the finding that HCMV-UL20, which is in close proximity to the HLA-like gene, encodes a protein with similarity to T-cell receptor γ-chains (BECK and BARRELL, unpublished observations). However, the match is marginal in nature, and alignment of a single region with both the constant (Cγ) and variable (Vγ) TCRγ-regions is possible. The former alignment shows approximately 16% identity over 194 amino acids, while the latter has approximately 27% identity over 82 amino acids. Although the Cγ alignment matches four cysteines, two on each side of the transmembrane domain, the remainder of the alignment is less convincing. In contrast, the Vγ alignment contains at least three localized clusters of homology including a highly conserved cysteine residue. However, a disulfide bond formed within Vγ may not be conserved; in HCMV-UL20 the second cysteine residue is located in the putative transmembrane domain. It is clear that no conclusions can be drawn regarding the significance of this match on the basis of the alignment. As in the case of the HLA homolog, sequence data from wild-type isolates might clarify the situation. If HCMV-UL20 is in fact a TCR homolog, the virus could exploit the interaction between TCRγ and CD3 to infect T cells, which might parallel the interaction of CD4 with the HIV gp120 protein (BORST et al. 1987; BRENNER et al. 1987). Furthermore, it is interesting to note that a feline retrovirus has been shown to encode a TCRβ-gene (FULTON et al. 1987).

7 Gene Families

In addition to gB and gH, several small glycoprotein genes were identified in HCMV, in US (WESTON and BARRELL 1986). These are arranged tandemly and tend to cluster as homologous blocks of reading frames, constituting a large proportion of the gene families found in HCMV. Interestingly, the HSV US glycoprotein genes are also clustered (DAVISON and MCGEOCH 1986; MCGEOCH et al. 1988a). We currently recognize nine sets of homologous genes in the AD169 genome. There are three pairs (UL25 and UL35; UL82 and UL83; and US2 and 3) and six larger groups. Of the latter, three occur in US where they account for a total of at least 21 genes (WESTON and BARRELL 1986); one family occurs in UL and RL; and two families are partitioned between the long and the short regions of the genome (Table 1). The discovery of redundant protein coding sequences outside repeat regions was unexpected and presents a contrast to those single genes encoding multiple products (for example, see Sects. 6.4 and 6.5). Their presence also appears to contradict the virally frugal gene layout of HCMV. As individual family members are likely to have subtle differences in function, this paradox may be difficult to resolve. The characteristics of four gene families are discussed below. Proteins have been recognized for three of these, while the fourth is homologous to a class of cellular receptors. The evolutionary implications of these findings are discussed in Sect. 8.

7.1 The RL11 Family

This family comprises fourteen members distributed in the long repeats and a portion of UL adjacent to TRL (Table 1; Fig. 1). The sequences are characterized by a motif which resembles the cellular Thy-1 in a region which is conserved with some other members of the immunoglobulin superfamily (C.A. HUTCHISON III, unpublished observations). The members of the RL11 Family are predicted to be membrane glycoproteins (Table 3). This prediction has been substantiated by the immunological detection of the Towne UL4-equivalent protein in infected cells and virions (CHANG et al. 1989a). The detected 48 kd protein is expressed during the early phase of infection, and its presence in virions has led to its classification as an early structural glycoprotein (CHANG et al. 1989a). Its published amino acid sequence is 84% identical to UL4 over 150 amino acids. Multiple alignment of the RL11 family suggests that UL4 (which does not contain an anchor sequence) may be spliced to UL5 (which has an anchor but no signal or N-glycosylation sites), as their respective RL11 homologous regions appear to dovetail somewhat. However, splicing was not observed in transcript mapping experiments (CHANG et al. 1989b). Nevertheless, CHANG et al. (1989a) detect a protein reduced in size from 48 kd to 27 kd protein when infected cells are treated with an inhibitor of N-linked glycosylation, although the theoretical size of UL4 alone is approximately 17 kd. While this difference could be attributable to other post-translational modifications, it is noteworthy that the theoretical size of RL11, which is homologous to both UL4 and UL5, is approximately 27 kd. The mapped transcripts, which are initiated from

three different promoters, also contain the UL5 reading frame. Hence it may be of interest to further characterize the 27 kDA protein. UL8 is truncated similarly to UL5, and therefore is also a candidate for splicing. As both these frames also contain KOZAK consensus ATG codons, a potential exists for the expression of this gene family to be regulated in a complex manner.

7.2 The US6 Family

This family corresponds to family 2 described by WESTON and BARRELL (1986) and is characterized by two areas of sequence homology, the second of which (region 2 (WESTON and BARRELL 1986)) is less well conserved. The region 1 core motif can be defined as C(VY)x(DQKR) (7–10) WxxxGxF where the bracketed residues are alternatives and x is any residue. The region 2 motif is characterized by cysteine and proline residues: PCxxC (4–6) CxPxxxxPWxP. The six members of this family are predicted to be membrane glycoproteins (Tables 1 and 3). GRETCH et al. (1988b) have recently used a MAb to demonstrate that this family correlates with the gp47–52 virion envelope glycoprotein complex they described previously (GRETCH et al. 1988a). Northern hybridization revealed three early transcripts from this region, two of which were minor species. The 1.6-kb size of the major transcript was consistent with initiation from the HCMV-US11 (HXLF1) TATA box, and in vitro translation experiments suggested it was bicistronic in nature. GRETCH et al. (1988a) suggest on the basis of these data and amino acid composition analysis that the main constituents of gp47–52 might be HCMV-US10 and US11 proteins. However, no direct correlation was established between the abundance of the putative transcript and the composition of gp47–52.

7.3 The US22 Family

This family is distributed in UL, US and RS and sequences for eight of the thirteen recognized members have been published, including the family 4 members described by WESTON and BARRELL (1986). Genes attributed to this family contain one or more of three sequence motifs (KOUZARIDES et al. 1988). The first motif (ooCCxxxLxxoG, where o is any hydrophobic residue and x any residue) is found in all of the members except IRS/TRS1 and UL28. Interestingly, in HCMV-UL36 the junction of exons 1 and 2 occurs immediately before the motif (KOUZARIDES et al. 1988). As HCMV-UL42 ends within the motif (FLCCDKFLPG- COO^-), it seems possible that this gene, and perhaps other members of the family apart from HCMV-UL36, encode spliced transcripts. The remainder of the pattern comprises two motifs which are largely hydrophobic and may overlap in function. The IRS/TRS1 genes, identical over most of their length, diverge shortly after the third motif. Apart from the conserved motifs, several of these sequences contain short runs of charged residues in their carboxy-terminal domains, and 6 of the 12 members of the US22 gene family have at least 1 N-linked glycosylation site. However, there does not appear to be any obvious correlation between these latter features. The only present correlation

between this gene family and viral proteins comes from the identification of the HCMV-US22 gene product ICP22. This is an early protein localizing in the nucleus which is also detectable in the cytoplasm and may be secreted from infected cells (MOCARSKI et al. 1988). Interestingly, the MAb used identifying US22 does not appear to recognize any of its homologs.

7.4 The G-Protein Coupled Receptor (GCR) Family

Several HCMV-reading frames, mostly located in US, are predicted to be integral membrane proteins capable of spanning the membrane several times (Table 1). All of these have seven potential membrane-spanning regions. Three of the reading frames, HCMV-US27 and HCMV-US28 (originally named HHRF2 and HHRF3; WESTON and BARRELL 1986), and HCMV-UL33, show homology to the opsin family of cell surface receptors (CHEE et al., submitted). Members of this diverse family of receptors

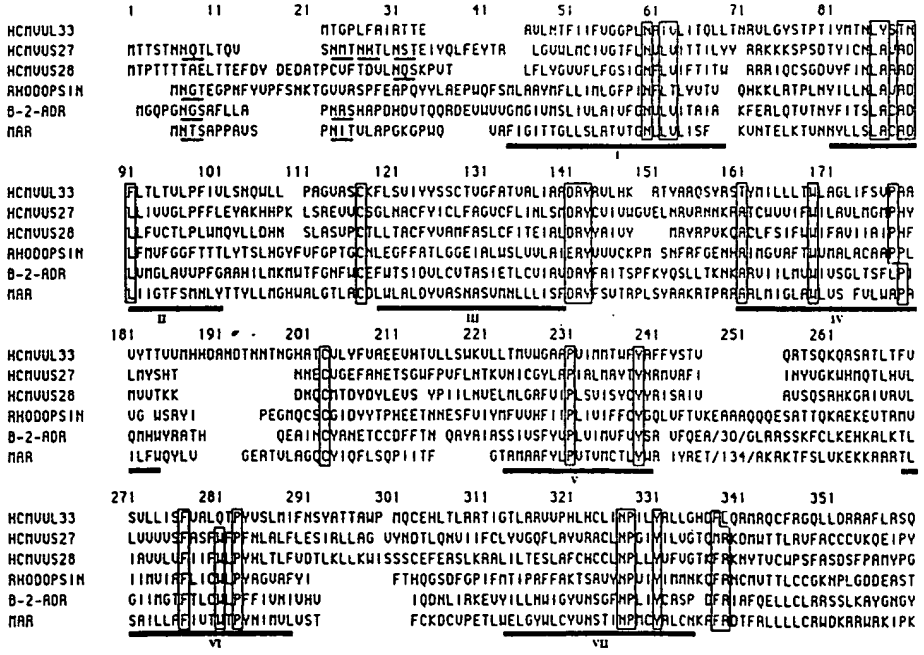


Fig. 4. An alignment of the three HCMV G-protein-coupled receptor homologs with bovine rhodopsin (NATHANS and HOGNESS 1983), human β-2-adrenergic receptor (B-2-ADR) (KOBILKA et al. 1987), and porcine muscarinic acetylcholine receptor (MAR) (KUBO et al. 1986). The NXT/S motifs are *underlined* in the N-terminal extracellular domain and identities which correspond in at least five of the six sequences are *boxed*. The seven membrane-spanning helical domains are indicated by *numbered bars* beneath the alignment. Each transmembrane domain and its disposition is defined by a motif unique within the sequence. The alignment has been truncated within the cytoplasmic C-terminal domains which possess receptor-specific functions, and sections of 30 and 134 amino acids have been excised from the B-2-ADR and MAR sequences respectively beginning at position 248. The two conserved cysteine residues at alignment positions 117 and 203 have been shown to be essential for function in bovine rhodopsin (KARNIK et al. 1988)

transduce different signals in a variety of systems, and have roles in vision, olfaction, memory and learning, and regulation of the circulatory system, among others (DOHLMANN et al. 1987; NATHANS 1987). The best-known subgroups of this family are the rhodopsins which absorb light via bound 11-*cis*-retinal, the β-adrenergic receptors which binds catecholamine hormones, and the muscarinic acetylcholine receptors. All of the above transduce signals through the membrane by activating G proteins. HCMV-US27, US28, and UL33 show the same membrane-spanning topography, are of similar size (362, 323, and 390 amino acids respectively), and are probably unspliced. US27 and US28 also have N-linked glycosylation sites at the N-terminus in common with the cellular members of the family. Apart from the overall similarity there is homology at the amino acid level mostly in and around the membrane-spanning sequences. An alignment of these sequences is shown in Fig. 4. The homology consists of short motifs that can uniquely define each membrane-spanning segment. At present the function of these genes is unknown. However, the downstream signal amplification by many of these receptors involves cAMP synthesis, which is suggestive in light of the presence of cAMP-responsive elements in the major immediate-early gene enhancer (Sect. 5.1).

8 Relationships to α and γ-Herpesvirus Genomes

The accumulated sequence data have begun to provide a broad evolutionary view of the herpesvirus family as a whole (HONESS 1984; HONESS et al. 1989). One feature in the evolution of herpesviruses is the movement of gene blocks within the genome, resulting in new arrangements of genes and presumably the disruption and formation of genes at recombinatorial junctions. Figure 5 shows the relationships of conserved sequences between the long unique regions of the sequenced human herpesviruses. The relationships between these regions of VZV, EBV, and HSV-1 have been analyzed previously (DAVISON and TAYLOR 1987; MCGEOCH et al. 1988a; MCGEOCH 1987). A comparison of the gene layout in HCMV *Hin*dIII F to equivalent regions in EBV and HSV-1 has also been published (KOUZARIDES et al. 1987b). As can be seen from Fig. 5, while the gene layouts of EBV and the α-herpesviruses are grossly more similar to each other than to HCMV, there do not appear to be any large blocks of genes that are not conserved between all three of the herpesvirus families. This is consistent with the notion that a core of herpesvirus genes is common to, and helps to define, the herpesvirus type. It also suggests that the three families of herpesviruses have diverged to such an extent that at the genetic level little else than this core set of genes remains in common between them. However, at the protein sequence level HCMV is more closely related to EBV than the α-herpesviruses, while the genes within each block show widely varying levels of conservation, ranging to undetectable or nonexistent (Table 2). While sequence comparisons with other herpesviruses help in establishing cladistic relationships, the following distinctive features of the HCMV genome give additional clues to its evolutionary past:

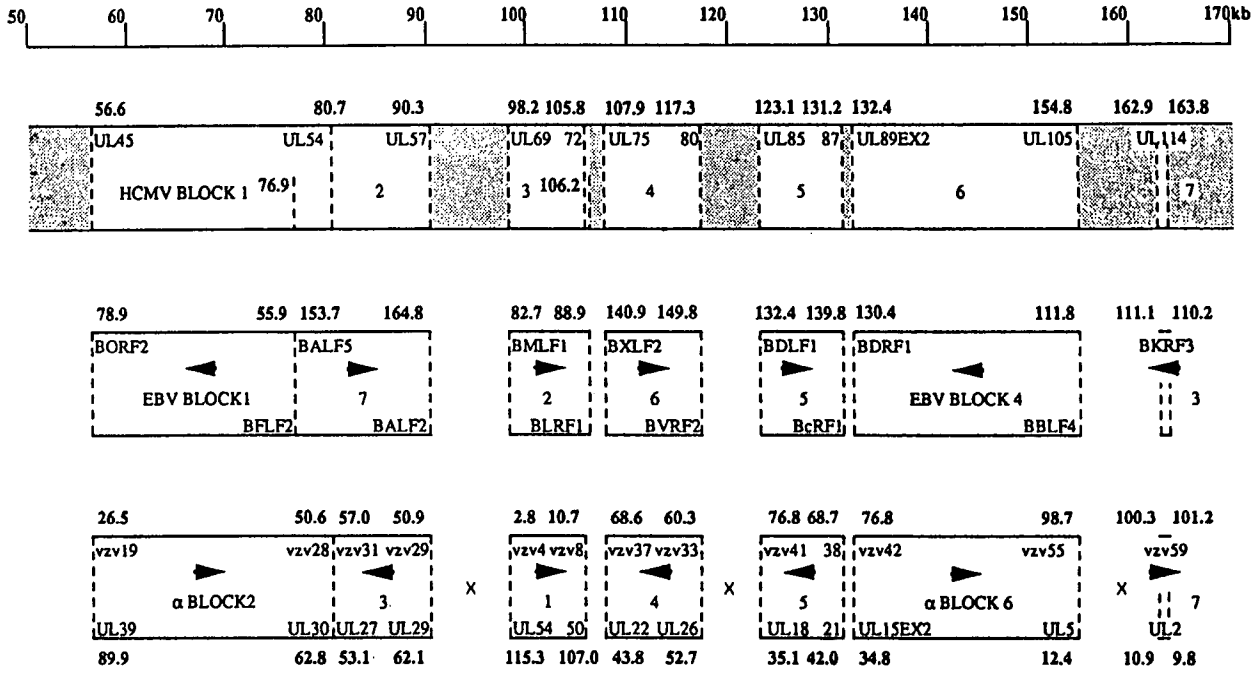


Fig. 5. Conserved blocks of sequence between HCMV and EBV, VZV, and HSV-1. The *uppermost map* represents a section of the HCMV UL indicated by the scale at the top of the diagram. The *middle map* depicts regions of EBV conserved with HCMV, and the *lower map* shows VZV and HSV-1 homologies, also to HCMV. Only the HCMV map is drawn to scale. All homologies found so far with the α- and γ-herpesviruses are located within the *unshaded* sections of the HCMV UL. The approximate boundary positions of the homology blocks within their respective genomes are marked in *boldtype* in kilobase pairs (positions are taken from Table 1 and BAER et al. 1984; DAVISON and SCOTT 1986; MCGEOCH et al. 1988a). Note that these numbers represent only the termini of the endmost detected homologous frames in each genome, and that some of these homologies are tentative (Table 2). The names of the frames are given. The orientation of each of the blocks in EBV and VZV (but not HSV-1) is shown relative to their published maps (BEAR et al. 1984; DAVISON and SCOTT 1986); *rightward arrowheads* denote collinearity. The order of the blocks within each genome is shown by a *block number*; these read from left to right across the genome in ascending order. Three of the five locations of nonhomologous reading frames found between the UL regions of HSV-1 and VZV are marked in the lower map (x) (MCGEOCH et al. 1988a)

The genes in HCMV that are conserved in the other herpesvirus families all appear to lie between approximately 50 to 170 kb in UL on the prototype genome. In contrast the extended HCMV gene families and the majority of the glycoprotein genes lie within US and in UL at left hand end of the prototype genome. Members of two families (the RL11 and US22 families) occur in RL and RS.

Two families (the US22 and GCR families) are partitioned between the short and the long regions of the genome. It also seems possible that the RL11, US2, and US6 families, together with HCMV-US34, are all members of a HCMV gene "superfamily" which is also partitioned between the short and long regions. These sequences all encode glycoproteins (or putative glycoprotein exons) which are mostly in the range of 200 amino acids in length. Multiple sequence alignment reveals short regions of amino acid homology between US2 and US3 and some members of the RL11 family. For example, the RL11 family anchor sequences are characterized by the motif HxxW, which is also seen in US2 (Table 3). The distinguishing motifs of the RL11 and US6 families also show some similarity, and may also be echoed in HCMV-US34:

RL11 Family motif:	Cxx (QEKR)	(7–10)	W		xxx	GxF
US6 Family motif:	Cxx (NQEKTY)	(4–6)	(YFLI)	Nx (ST)	xxxx	GxY
HCMV-US34:	CLAE	VGVA		NAT	FLSRFNV	GDF

Finally, the majority of the genes in families are present as tandemly repeated copies. These observations suggest that the HCMV gene repertoire has been expanding by gene duplication and divergence, a process which may be mediated by the HCMV DNA replication machinery (WEBER et al. 1988) and which may be related to expansion and contraction of repeat sequences (WHITTON and CLEMENTS 1984; DAVISON and MCGEOCH 1986). Furthermore, there appears to have been at least one recombination event involving the long and short regions of HCMV which led to the distribution of gene families between both regions. A possible scenario for such an event might be an internal duplication of a terminal segment leading to the conversion of an ancestral non-inverting genome to a four-isomer genome. Genes partitioned between the repeats of the two new subsegments might then diverge, together with the expansion and contraction of the repeats. The characterization of other betaherpesvirus sequences may help to clarify the evolutionary history of HCMV, and it will be of interest to see if the propensity of HCMV for gene duplication is a general charcteristic of the β-herpesviruses.

9 Perspectives

This project is a contribution to a set of genomic sequences which now represents the three main branches of the herpesvirus family. The prior sequencing of EBV, VZV, and HSV-1 has greatly facilitated the analysis of the HCMV genome, and features

which unify this highly divergent group of viruses are now coming into focus at the genetic level. The sequences have facilitated the correlation of biological and genetic experiments, and allowed much of this work to be generalized. The growing body of relational knowledge should make it increasingly informative to begin the characterization of herpesvirus genomes by sequencing. These data will continue to provide predictions which can be tested, and which promise to shed further light on the herpesviruses and their eukaryotic environment.

Acknowledgments. We thank Jon Oram and Peter Greenaway for providing the *Hin*dIII clones used in the sequence analysis and Bernard Fleckenstein for providing the cosmid clones used for determining the overlaps of the *Hin*dIII sites. We are grateful to Tony Minson and Helena Browne for advice and for making available results prior to publication, and to Mark Stinski for comments on parts of the manuscript. M.C. thanks the Commonwealth Scholarships Commission for support.

References

Addison C, Rixon FJ, Palfreyman JW, O'Hara M, Preston VG (1984) Characterisation of a herpes simplex virus type 1 mutant which has a temperature-sensitive defect in penetration of cells and assembly of capsids. Virology 138: 246–259

Akrigg A, Wilkinson GWG, Oram JD (1985) The structure of the major immediate early gene of human cytomegalovirus strain AD169. Virus Res 2: 107–121

Anders DG, Gibson W (1988) Location, transcript analysis, and partial nucleotide sequence of the cytomegalovirus gene encoding an early DNA-binding protein with similarities to ICP8 of herpes simplex virus type 1. J Virol 62: 1364–1372

Avertt DR, Lubbers C, Elion GB, Spector T (1983) Ribonucleotide reductase induced by herpes simplex virus type 1. Characterisation of a distinct enzyme. J Biol Chem 258: 9831–9838

Baer R, Bankier AT, Biggin MD, Deininger PL, Farrell PJ, Gibson TJ, Hatfull G, et al. (1984) DNA sequence and expression of the B95-8 Epstein-Barr virus genome. Nature 310: 207–211

Bairoch A (1988) Swiss-Prot protein sequence data bank release 8.0. Department de Biochimie Medicale, Centre Medical Universitaire, Geneva

Bankier AT, Barrell BG (1989) Sequencing single strand DNA using the chain termination method. In: Ward S, Howe C (eds) Nucleic acids sequencing: a practical approach. IRL, Oxford (in press)

Bankier AT, Weston KM, Barrell BG (1988) Random cloning and sequencing by the M13/dideoxynucleotide chain termination method. Methods Enzymol 155: 51–93

Batterson W, Furlong D, Roizman B (1983) Molecular genetics of herpes simplex virus VIII. Further characterization of a temperature-sensitive mutant defective in release of viral DNA and in other stages of the viral reproductive cycle. J Virol 45: 397–407

Beck S, Barrell BG (1988) Human cytomegalovirus encodes a glycoprotein homologous to MHC class-I antigens. Nature 331: 269–272

Benko DM, Haltiwanger RS, Hart GW, Gibson W (1988) Virion basic phosphoprotein from human cytomegalovirus contains *O*-linked *N*-acetyl glucosamine. Proc Natl Acad Sci USA 85: 2573–2577

Biron KK, Fyfe JA, Stanat SC, Leslie LK, Sorrell JB, Lambe CU, Coen DM (1986) A human cytomegalovirus mutant resistant to the nucleoside analog 9-{[2-hydroxy-1-(hydroxymethyl)ethoxy]methyl} guanine (BW B759U) induces reduced levels of BW B759U triphosphate. Proc Natl Acad Sci USA 83: 8769–8773

Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC (1987a) Structure of the human class I histocompatibility antigen, HLA-A2. Nature 329: 506–512

Bjorkman PJ, Saper MA, Samraoui B, Bennett WS, Strominger JL, Wiley DC (1987b) The foreign antigen binding site and T-cell recognition regions of class I histocompatibility antigens. Nature 329: 512–518

Borst J, van de Griend RJ, van Oostveen JW, Ang S-L, Melief CJ, Seidman JG, Bolhuis RLH (1987) A T-cell receptor γ/CD3 complex found on cloned functional lymphocytes. Nature 325: 683–688
Boshart M, Weber F, Jahn G, Dorsch-Hasler K, Fleckenstein B, Schaffner W (1985) A very strong enhancer is located upstream of an immediate early gene of human cytomegalovirus. Cell 41: 521–530
Brenner S (1987) Phosphotransferase sequence homology. Nature 329: 21
Brenner MB, McLean J, Scheft H, Riberdy J, Ang S-L, Seidman JG, Devlin P, Krangel MS (1987) Two forms of the T-cell receptor γprotein found on peripheral blood cytotoxic T lymphocytes. Nature 325: 689–694
Chang C-P, Vesole DH, Nelson J, Oldstone MBA, Stinski MF (1989a) Identification and expression of a human cytomegalovirus early glycoprotein. J Virol 63: 3330–3337
Chang C-P, Malone CL, Stinski MF (1989b) A human cytomegalovirus early gene has three inducible promoters that are regulated differentially at various times after infection. J Virol 63: 281–290
Chee MS, Lawrence GL, Barrell BG (1989a) Alpha-, beta-, and gammaherpesviruses encode a putative phosphotransferase. J Gen Virol 70 (in press)
Chee MS, Rudolph S-A, Plachter B, Barrell BG, Jahn G (1989b) Identification of the major capsid protein gene of human cytomegalovirus. J Virol 63: 1345–1353
Chee MS, Satchwell SC, Preddie E, Weston KM, Barrell BG. Human cytomegalovirus encodes three G-protein coupled receptor homologues. Submitted for publication.
Cherrington JM, Mocarski ES (1989) Human cytomegalovirus ie1 transactivates the α promoter-enhancer via an 18-base-pair repeat element. J Virol 63: 1435–1440
Chou J, Roizman B (1989) Characterization of DNA sequence-common and sequence-specific proteins binding to cis-acting sites for cleavage of the terminal *a* sequence of the herpes simplex virus 1 genome. J Virol 63: 1059–1068
Clark BR, Zaia JA, Balce-Directo L, Ting Y-P (1984) Isolation and partial chemical characterization of a 64,000-dalton glycoprotein of human cytomegalovirus. J Virol 49: 279–282
Costa RH, Draper KG, Kelly TJ, Wagner EK (1985) An unusual spliced herpes simplex virus type 1 transcript with sequence homology to Epstein-Barr virus DNA. J Virol 54: 317–328
Cranage MP, Kouzarides T, Bankier AT, Satchwell SC, Weston KW, Tomlinson P, Barrell BG, et al. (1986) Identification of the human cytomegalovirus glycoprotein B gene and induction of neutralizing antibodies via its expression in recombinant vaccinia virus. EMBO J 5: 3057–3063
Cranage MP, Smith GL, Bell SE, Hart H, Brown C, Bankier AT, Tomlinson P, et al. (1988) Identification and expression of a human cytomegalovirus glycoprotein with homology to the Epstein-Barr virus BXLF2 product, varicella-zoster virus gpIII, and herpes simplex virus type 1 glycoprotein H. J Virol 62: 1416–1422
Crute JJ, Mocarski ES, Lehman IE (1988) A DNA helicase induced by herpes simplex virus type 1. Nucleic Acids Res 16: 6585–6596
Crute JJ, Tsurumi T, Zhu L, Weller SK, Olivo PD, Challberg MD, Mocarski ES, Lehman IR (1989) Herpes simplex virus 1 helicase-primase: a complex of three herpes-encoded gene products. Proc. Natl Acad Sci USA 86: 2186–2189
Davis MG, Huang E-S (1985) Nucleotide sequence of a human cytomegalovirus DNA fragment encoding a 67-kilodalton phosphorylated viral protein. J Virol 56: 7–11
Davis MG, Mar E-C, Wu Y-M, Huang E-S (1984) Mapping and expression of a human cytomegalovirus major viral protein. J Virol 52: 129–135
Davison AJ, McGeoch DJ (1986) Evolutionary comparisons of the S segments in the genomes of herpes simplex virus type 1 and varicella-zoster virus. J Gen Virol 67: 597–611
Davison AJ, Scott JE (1986) The complete DNA sequence of varicella-zoster virus. J Gen Virol 67: 1759–1816
Davison AJ, Taylor P (1987) Genetic relations between varicella-zoster virus and Epstein-Barr virus. J Gen Virol 68: 1067–1079
Del Val M, Munch K, Reddehase MJ, Koszinowski UH (1989) Presentation of CMV immediate-early antigen to cytolytic T lymphocytes is selectively prevented by viral genes expressed in the early phase. Cell 58: 305–315
DeMarchi JM (1981) Human cytomegalovirus DNA: restriction enzyme cleavage maps and map locations for immediate-early, early, and late RNAs. Virology 114: 23–38
DeMarchi JM (1983) Post-transcriptional control of human cytomegalovirus gene expression. Virology 124: 390–402
Depto AS, Stenberg RM (1989) Regulated expression of the human cytomegalovirus pp65 gene: octamer sequence in the promoter is required for activation by viral gene products. J Virol 63: 1232–1238
Desai PJ, Schaffer PA, Minson AC (1988) Excretion of non-infectious virus particles lacking glycoprotein H by a temperature-sensitive mutant of herpes simplex virus type 1: evidence that gH is essential for virion infectivity. J Gen Virol 69: 1147–1156

Dohlman HG, Caron MG, Lefkowitz RJ (1987) A family of receptors coupled to guanine nucleotide regulatory proteins. Biochemistry 26: 2657–2664
Dorsch-Hasler K, Keil GM, Weber F, Jasin M, Schaffner W, Koszinowski UH (1985) A long and complex enhancer activates transcription of the gene coding for the highly abundant immediate early mRNA in murine cytomegalovirus. Proc Natl Acad Sci USA 82: 8325–8329
Engstrom Y, Francke U (1985) Assignment of the structural gene for subunit M1 of human ribonucleotide reductase to the short arm of chromosome 11. Exp Cell Res 158: 477–483
Farrar GH, Greenaway PJ (1986) Characterization of glycoprotein complexes present in human cytomegalovirus envelopes. J Gen Virol 67: 1469–1473
Ferguson MAJ, Williams AF (1988) Cell-surface anchoring of proteins via glycosyl-phosphatidylinositol structures. Annu Rev Biochem 57: 285–320
Fickenscher H, Stamminger T, Ruger R, Fleckenstein B (1989) The role of a repetitive palindromic sequence element in the human cytomegalovirus immediate early enhancer. J Gen Virol 70: 107–123
Fleckenstein B, Muller I, Collins J (1982) Cloning of the complete human cytomegalovirus genome in cosmids. Gene 18: 39–46
Fulton R, Forrest D, McFarlane R, Onions D, Neil JC (1987) Retroviral transduction of T-cell antigen receptor β-chain and *myc* genes. Nature 326: 190–194
Geballe AP, Mocarski ES (1988) Translational control of cytomegalovirus gene expression is mediated by upstream AUG codons. J Virol 62: 3334–3340
Geballe AP, Spaete RR, Mocarski ES (1986a) A *cis*-acting element within the 5′ leader of a cytomegalovirus β transcript determines kinetic class. Cell 46: 865–872
Geballe AP, Leach FS, Mocarski EM (1986b) Regulation of cytomegalovirus late gene expression: γ genes are controlled by posttranscriptional events. J Virol 57: 864–874
George DG, Barker WC, Hunt LT (1986) The protein identification resource (PIR). Nucleic Acids Res 14: 11–15
Ghazal P, Lubon H, Fleckenstein B, Hennighausen L (1987) Binding of transcription factors and creation of a large nucleoprotein complex on the human cytomegalovirus enhancer. Proc Natl Acad Sci USA 84: 3658–3662
Ghazal P, Lubon H, Hennighausen L (1988) Specific interactions between transcription factors and the promoter-regulatory region of the human cytomegalovirus major immediate-early gene. J Virol 62: 1076–1079
Gibson W (1983) Protein counterparts of human and simian cytomegalovirus. Virology 128: 391–406
Gibson T, Stockwell P, Ginsburg M, Barrell BG (1984) Homology between two EBV early genes and HSV ribonucleotide reductase and 38K genes. Nucleic Acids Res 12: 5087–5099
Goins WF, Stinski MF (1986) Expression of a human cytomegalovirus late gene is posttranscriptionally regulated by a 3′-end-processing event occurring exclusively late after infection. Mol Cell Biol 6: 4202–4213
Gompels UA, Craxton MA, Honess RW (1988a) Conservation of gene organization in the lymphotropic herpesviruses herpesvirus saimiri and Epstein-Barr virus. J Virol 62: 757–767
Gompels UA, Craxton MA, Honess RW (1988b) Conservation of glycoprotein H (gH) in herpesviruses: nucleotide sequence of the gH gene from herpesvirus saimiri. J Gen Virol 69: 2819–2829
Greenaway PJ, Wilkinson GWG (1987) Nucleotide sequence of the most abundantly transcribed early gene of human cytomegalovirus strain AD169. Virus Res 7: 17–31
Gretch DR, Kari B, Rasmussen L, Gehrz RC, Stinski MF (1988a) Identification and characterization of three distinct families of glycoprotein complexes in the envelopes of human cytomegalovirus. J Virol 62: 875–881
Gretch DR, Kari B, Gehrz RC, Stinski MF (1988b) A multigene family encodes the human cytomegalovirus glycoprotein complex gcII (gp47–52 complex). J Virol 62: 1956–1962
Grundy JE, McKeating JA, Griffiths PD (1987a) Cytomegalovirus strain AD169 binds β_2 microglobulin in vitro after release from cells. J Gen Virol 68: 777–784
Grundy JE, McKeating JA, Ward PJ, Sanderson AR, Griffiths PD (1987b) β_2 Microglobulin enhances the infectivity of cytomegalovirus and when bound to the virus enables class 1 HLA molecules to be used as a virus receptor. J Gen Virol 68: 793–803
Heilbronn R, Jahn G, Burkle A, Freese U-K, Fleckenstein B, zur Hausen H (1987) Genomic localization, sequence analysis, and transcription of the putative human cytomegalovirus DNA polymerase gene. J Virol 61: 119–124
Hennighausen L, Fleckenstein B (1986) Nuclear factor 1 interacts with five DNA elements in the promoter region of the human cytomegalovirus major immediate early gene. EMBO J 5: 1367–1371
Hermiston TW, Malone CL, Witte PR, Stinski MF (1987) Identification and characterization of the human cytomegalovirus immediate-early region 2 gene that stimulates gene expression from an inducible promoter. J Virol. 61: 3214–3221
Hodgman TC (1988) A new superfamily of replicative proteins. Nature 333: 22–23

Honess RW (1984) Herpes simplex and the 'herpes complex': diverse observations and a unifying hypothesis. J Gen Virol 65: 2077–2107
Honess RW, Bodemer W, Cameron KR, Niller H-H, Fleckenstein B, Randall RE (1986) The A + T-rich genome of herpesvirus saimiri contains a highly conserved gene for thymidylate synthase. Proc Natl Acad Sci USA 83: 3604–3608
Honess RW, Gompels UA, Barrell BG, Craxton M, Cameron KR, Staden R, Chang Y-N, Hayward GS (1989) Deviations from expected frequencies of CpG dinucleotides in herpesvirus DNAs may be diagnostic of differences in the states of their latent genomes. J Gen Virol 70: 837–855
Hunninghake GW, Monick MM, Liu B, Stinski MF (1989) The promoter-regulatory region of the major immediate-early gene of human cytomegalovirus responds to T-lymphocyte stimulation and contains functional cyclic AMP-response elements. J Virol 63: 3026–3033
Hutchinson NI, Tocci MJ (1986) Characterization of a major early gene from the human cytomegalovirus long inverted repeat; predicted amino acid sequence of a 30-kDa protein encoded by the 1.2 kb mRNA. Virology 155: 172–182
Hutchinson NI, Sondermeyer RT, Tocci MJ (1986) Organization and expression of the major genes from the long inverted repeat of the human cytomegalovirus genome. Virology 155: 160–171
Irmiere A, Gibson W (1983) Isolation and characterization of a noninfectious virion-like particle released from cells infected with human strains of cytomegalovirus. Virology 130: 118–133
Irmiere A, Gibson W (1985) Isolation of human cytomegalovirus intranuclear capsids, characterization of their protein constituents, and demonstration that the B-capsid assembly protein is also abundant in noninfectious enveloped particles. J Virol 56: 277–283
Jahan N, Razzaque A, Brady J, Rosenthal LJ (1989) The human cytomegalovirus mtrII colinear region in strain Tanaka is transformation defective. J Virol 63: 2866–2869
Jahn G, Knust E, Schmolla H, Sarre T, Nelson JA, McDougall JK, Fleckenstein B (1984) Predominant immediate-early transcripts of human cytomegalovirus AD169. J Virol 49: 363–370
Jahn G, Kouzarides T, Mach M, Scholl, B-C, Plachter B, Traupe B, Preddie E, et al. (1987) Map position and nucleotide sequence of the gene for the large structural phosphoprotein of human cytomegalovirus. J Virol 61: 1358–1367
Jeang K-T, Hayward GS (1983) A cytomegalovirus DNA sequence containing tracts of tandemly repeated CA dinucleotides hybridises to highly repetitive dispersed elements in mammalian cell genomes. Mol Cell Biol 3: 1389–1402
Jeang KT, Rawlins DR, Rosenfeld P, Shero JH, Kelly T, Hayward GS (1987) Multiple tandemly repeated binding sites for cellular nuclear factor 1 that surround the major immediate-early promoters of simian and human cytomegalovirus. J Virol 61: 1559–1570
Karnik SS, Sakmar TP, Chen H-B, Khorana HG (1988) Cysteine residues 110 and 187 are essential for the formation of correct structure in bovine rhodopsin. Proc Natl Acad Sci USA 85: 8459–8463
Keil GM, Ebeling-Keil A, Koszinowski UH (1987) Sequence and structural organization of murine cytomegalovirus immediate-early gene 1. J Virol 61: 1901–1908
Kobilka BK, Dixon RAF, Frielle T, Dohlman HG, Bolanowski MA, Sigal IS, Yang-Feng TL, et al. (1987) cDNA for the human β_2-adrenergic receptor: a protein with multiple membrane-spanning domains and encoded by a gene whose chromosomal location is shared with that of the receptor for platelet-derived growth factor. Proc Natl Acad Sci USA 84: 46–50
Kouzarides T, Bankier AT, Barrell BG (1983) Nucleotide sequence of the transforming region of human cytomegalovirus. Mol Biol Med 1: 47–58
Kouzarides T, Bankier AT, Satchwell SC, Weston K, Tomlinson P, Barrell BG (1987a) Sequence and transcription analysis of the human cytomegalovirus DNA polymerase gene. J Virol 61: 125–133
Kouzarides T, Bankier AT, Satchwell SC, Weston K, Tomlinson P, Barrell BG (1987b) Large-scale rearrangement of homologous regions in the genomes of HCMV and EBV. Virology 157: 397–413
Kouzarides T, Bankier AT, Satchwell SC, Preddie E, Barrell BG (1988) An immediate early gene of human cytomegalovirus encodes a potential membrane glycoprotein. Virology 165: 151–164
Kozak M (1981) Possible role of flanking nucleotides in recognition of the AUG initiator codon by eukaryotic ribosomes. Nucleic Acids Res 9: 5233–5252
Kozak M (1982) Analysis of ribosome binding sites from the s1 message of reovirus: initiation at the first and second AUG codons. J Mol Biol 156: 807–820
Kubo T, Fukuda K, Mikami A, Maeda A, Takahashi H, Mishina M, Haga T, et al. (1986) Cloning, sequencing and expression of complementary DNA encoding the muscarinic acetylcholine receptor. Nature 323: 411–416
Landini M-P, Michelson S (1988) Human cytomegalovirus proteins. Prog Med Virol 35: 152–185
Laniken H, Graslund A, Thelander L (1982) Induction of a new ribonucleotide reductase activity after infection of mouse L cells with pseudorabies virus. J Virol 41: 893–900
Leach FS, Mocarski ES (1989) Regulation of cytomegalovirus late-gene expression: differential use of three start sites in the transcriptional activation of ICP36 gene expression. J Virol 63: 1783–1791

Lee JY, Irmiere A, Gibson W (1988) Primate cytomegalovirus assembly: evidence that DNA packaging occurs subsequent to B capsid assembly. Virology 167: 87–96
Littler E, Zeuthen J, McBride AA, Trost-Sorensen E, Powell KL, Walsh-Arrand JE, Arrand JR (1986) Identification of an Epstein-Barr virus-coded thymidine kinase. EMBO J 5: 1959–1966
Mach M, Utz U, Fleckenstein B (1986) Mapping of the major glycoprotein gene of human cytomegalovirus. J Gen Virol 67: 1461–1467
Marschalek R, Amon-Bohm E, Stoerker J, Klages S, Fleckenstein B, Dingermann T (1989) CMER, an RNA encoded by human cytomegalovirus is most likely transcribed by RNA polymerase III. Nucleic Acids Res 17: 631–643
Martignetti JA (1987) Sequence analysis of HCMV. Dissertation, Cambridge University
Martinez J, St Jeor SC (1986) Molecular cloning and analysis of three cDNA clones homologous to human cytomegalovirus RNAs present during late infection. J Virol 60: 531–538
Martinez J, Lahijani RS, St Jeor SC (1989) Analysis of a region of the human cytomegalovirus (AD169) genome coding for a 25-kilodalton virion protein. J Virol 63: 233–241
McDonough SH, Spector DH (1983) Transcription in human fibroblasts permissively infected by human cytomegalovirus strain AD169. Virology 125: 31–46
McDonough SH, Staprans SI, Spector DH (1985) Analysis of the major transcripts encdoded by the long repeat of human cytomegalovirus strain AD169. J Virol 53: 711–718
McGeoch DJ (1985) On the predictive recognition of signal peptide sequences. Virus Res 3: 271–286
McGeoch DJ (1987) The genome of herpes simplex virus: structure, replication and evolution. J Cell Sci [Suppl] 7: 67–94
McGeoch DJ, Davison AJ (1986) Alphaherpesviruses possess a gene homologous to the protein kinase gene family of eukaryotes and retroviruses. Nucleic Acids Res 14: 1765–1777
McGeoch DJ, Dalrymple MA, Davison AJ, Dolan A, Frame MC, McNab D, Perry LJ, et al. (1988a) The complete sequence of the long unique region in the genome of herpes simplex virus type 1. J Gen Virol 69: 1531–1574
McGeoch DJ, Dolan A, Frame MC (1986) DNA sequence of the region in the genome of herpes simplex virus type 1 containing the exonuclease gene and neighbouring genes. Nucleic Acids Res 14: 3435–3448
McGeoch DJ, Dalrymple MA, Dolan A, McNab D, Perry L, Taylor P, Challberg MD (1988b) Structures of herpes simplex virus type 1 genes required for replication of virus DNA. J Virol 62: 444–453
McKnight SL (1980) The nucleotide sequence and transcript map of the herpes simplex virus thymidine kinase gene. Nucleic Acids Res 8: 5949–5963
Meyer H, Bankier AT, Landini MP, Brown CM, Barrell BG, Ruger B, Mach M (1988) Identification and procaryotic expression of the gene coding for the highly immunogenic 28-kilodalton structural phosphoprotein (pp28) of human cytomegalovirus. J Virol 62: 2243–2250
Mocarski ES, Roizman B (1982) Structure and role of the herpes simplex virus DNA termini in inversion, circularization and generation of virion DNA. Cell 31: 89–97
Mocarski ES, Pereira L, Michael N (1985) Precise localization of genes on large animal virus genomes: use of λgt11 and monoclonal antibodies to map the gene for a cytomegalovirus protein family. Proc Natl Acad Sci USA 82: 1266–1270
Mocarski ES, Pereira L, McCormick AL (1988) Human cytomegalovirus ICP22, the product of the HWLF1 reading frame, is an early nuclear protein that is released from cells. J Gen Virol 69: 2613–2621
Mullaney J, Moss HWMcL, McGeoch DJ (1989) Gene UL2 of herpes simplex virus type 1 encodes a uracil-DNA glycosylase. J Gen Virol 70: 449–454
Nathans J (1987) Molecular biology of visual pigments. Annu Rev Neurosci 10: 163–194
Nathans J, Hogness DS (1983) Isolation, sequence analysis, and intron-exon arrangement of the gene coding bovine rhodopsin. Cell 34: 807–814
Nikas I, McLauchlan J, Davison AJ, Taylor WR, Clements JB (1986) Structural features of ribonucleotide reductase. Proteins 1: 376–384
Olivo PD, Nelson NJ, Challberg MD (1988) Herpes simplex virus DNA replication: the UL9 gene encodes an origin-binding protein. Proc Natl Acad Sci USA 85: 5414–5418
Oram JD, Downing RG, Akrigg A, Doggleby CJ, Wilkinson GWG, Greenaway PJ (1982) Use of recombinant plasmids to investigate the structure of the human cytomegalovirus genome. J Gen Virol 59: 111–129
Pachl C, Probert WS, Hermsen KM, Masiarz FR, Rasmussen L, Merigan, TC, Spaete RR (1989) The human cytomegalovirus strain Towne glycoprotein H gene encodes glycoprotein p86. Virology 169: 418–426
Pande H, Baak SW, Riggs AD, Clark BR, Shively JE, Zaia JA (1984) Cloning and physical mapping of a gene fragment coding for a 64-kilodalton major late antigen of human cytomegalovirus. Proc Natl Acad Sci USA 81: 4965–4969

Pande H, Campo K, Churchill MA, Clark BR, Zaia JA (1988) Genomic localization of the gene encoding a 32-kDa capsid protein of human cytomegalovirus. Virology 167: 306–310

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85: 2444–2448

Pereira L, Hoffman M, Gallo D, Cremer N (1982) Monoclonal antibodies to human cytomegalovirus: three surface membrane proteins with unique immunological and electrophoretic properties specify cross-reactive determinants. Infect Immun 36: 924–932

Pertuiset B, Boccara M, Cerbrian J, Berthelot N, Chousterman S, Puvion-Dutilleul F, Sisman J, Sheldrick P (1989) Physical mapping and nucleotide sequence of a herpes simplex virus type 1 gene required for capsid assembly. J Virol 63: 2169–2179

Petrovskis EA, Timmins JG, Armentrout MA, Marchioli CC, Yancey RJ Jr, Post LE (1986) DNA sequence of the gene for pseudorabies virus gp50, a glycoprotein without N-linked glycosylation. J Virol 59: 216–223

Pizzorno MC, O'Hare P, Sha L, LaFemina RL, Hayward GS (1988) trans-Activation and autoregulation of gene expression by the immediate-early region 2 gene products of human cytomegalovirus. J Virol 62: 1167–1179

Preston VG, Fisher FB (1984) Identification of the herpes simplex virus type 1 gene encoding the dUTPase. Virology 138: 58–68

Preston VG, Coates JAV, Rixon FJ (1983) Identification and characterization of a herpes simplex virus gene product required for encapsidation of virus DNA. J Virol 45: 1056–1064

Rasmussen LE, Nelson RM, Kelsall DC, Merigan TC (1984) Murine monoclonal antibody to a single protein neutralizes the infectivity of human cytomegalovirus. Proc Natl Acad Sci USA 81: 876–880

Rasmussen RD, Staprans SI, Shaw SB, Spector DH (1985a) Sequences in human cytomegalovirus which hybridize with the avian retrovirus oncogene v-*myc* are G + C rich and do not hybridize with the human c-*myc* gene. Mol Cell Biol 5: 1525–1530

Rasmussen L, Mullenax J, Nelson R, Merigan TC (1985b) Viral polypeptides detected by a complement-dependent neutralizing murine monclonal antibody to human cytomegalovirus. J Virol 55: 274–280

Razzaque et al. (1988) Localization and DNA sequence analysis of the transforming domain (*mtrII*) of human cytomegalovirus. Proc Natl Acad Sci USA 85: 5709–5713

Reichard P (1989) Interactions between deoxyribonucleotide and DNA synthesis. Annu Rev Biochem 57: 349–374

Rixon FJ, Cross AM, Addison C, Preston VG (1988) The products of herpes simplex virus type 1 gene UL26 which are involved in DNA packaging are strongly associated with empty but not with full capsids. J Gen Virol 69: 2879–2891

Robson L, Gibson W (1989) Primate cytomegalovirus assembly protein: genome location and nucleotide sequence. J. Virol. 63: 669–676

Roby C, Gibson W (1986) Characterization of phosphoproteins and protein kinase activity of virions, noninfectious enveloped particles, and dense bodies of human cytomegalovirus. J Virol 59: 714–727

Ruger B, Klages S, Walla B, Albrecht J, Fleckenstein B, Tomlinson P, Barrell BG (1987) Primary structure and transcription of the genes coding for the two virion phosphoproteins pp65 and pp71 of human cytomegalovirus. J Virol 61: 446–453

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239: 487–491

Sjoberg B-M, Eklund H, Fuchs JA, Carlson J, Standart NM, Ruderman JV, Bray SJ, Hunt T (1985) Identification of the stable free radical tyrosine residue in ribonucleotide reductase. FEBS Lett 183: 99–102

Smith RF, Smith TF (1989) Identification of new protein kinase-related genes in three herpes viruses, herpes simplex virus, varicella-zoster virus and Epstein-Barr Virus. J Virol 63: 450–455

Spaete RR, Mocarski ES (1985a) Regulation of cytomegalovirus gene expression; α and β promoters are *trans* activated by viral functions in permissive human fibroblasts. J Virol 56: 135–143

Spaete RR, Mocarski ES (1985b) The a sequence of the cytomegalovirus genome functions as a cleavage/packaging signal for herpes simplex virus defective genomes. J Virol. 54: 817–824

Spaete RR, Mocarski ES (1987) Insertion and deletion mutagenesis of the human cytomegalovirus genome. Proc Natl Acad Sci USA 84: 7213–7217

Spaete RR, Thayer RM, Probert WS, Masiarz FR, Chamberlain SH, Rasmussen L, Merigan TC, Pachl C (1988) Human cytomegalovirus strain Towne glycoprotein B is processed by proteolytic cleavage. Virology 167: 207–225

Staden R (1986) The current status and portability of our sequencing handling software. Nucleic Acids Res 14: 217–231

Staden R (1988) Methods to define and locate patterns of motifs in sequences. CABIOS 4: 53–60

Stannard LM (1989) β_2 microglobulin binds to the tegument of cytomegalovirus: an immunogold study. J Gen Virol 70: 2179–2184
Staprans SI, Spector DH (1986) 2.2-kilobase class of early transcripts encoded by cell-related sequences in human cytomegalovirus strain AD169. J Virol 57: 591–602
Stenberg RM, Thomsen DR, Stinski MF (1984) Structural analysis of the major immediate early gene of human cytomegalovirus. J Virol 49: 190–191
Stenberg RM, Witte PR, Stinski MF (1985) Multiple spliced and unspliced transcripts from human cytomegalovirus immediate-early region 2 and evidence for a common initiation site within immediate-early region 1. J Virol 56: 665–675
Stinski MF (1977) Synthesis of proteins and glycoproteins in cells infected with human cytomegalovirus. J Virol 23: 751–767
Stinski MF, Roehr TJ (1985) Activation of the major immediate early gene of human cytomegalovirus by *cis*-acting elements in the promoter-regulatory sequence and by virus-specific *trans*-acting components. J Virol 55: 431–441
Stinski MF, Thomsen DR, Stenberg RM, Goldstein LC (1983) Organization and expression of the immediate early genes of human cytomegalovirus. J Virol 46: 1–14
Tamashiro JC, Filpula D, Friedmann T, Spector DH (1984) Structure of the heterogeneous L-S junction region of human cytomegalovirus strain AD169 DNA. J Virol 52: 541–584
Thompson R, Honess RW, Taylor L, Morran J, Davison AJ (1987) Varicella-zoster virus specifies a thymidylate synthetase. J Gen Virol 68: 1449–1455
Thomsen DR, Stenberg RM, Goins WF, Stinski MF (1984) Promoter-regulatory region of the major immediate early gene of human cytomegalovirus. Proc Natl Acad Sci USA 81: 659–663
Townsend A, Ohlen C, Bastin J, Ljunggren H-G, Foster L, Karre K (1989) Association of class I major histocompatibility heavy and light chains induced by viral peptides. Nature 340: 443–448
Trimble JJ, Murthy CS, Bakker A, Grassmann R, Desrosiers RC (1988) A gene for dihydrofolate reductase in a herpesvirus. Science 239: 1145–1147
Wang F, Petti L, Braun D, Seung S, Kieff E (1987) A bicistronic Epstein-Barr virus mRNA encodes two nuclear proteins in latently infected, growth-transformed lymphocytes. J Virol 61: 945–954
Wathen MW, Stinski MF (1982) Temporal patterns of human cytomegalovirus transcription: mapping the viral RNAs synthesized at immediate early, early, and late times after infection. J Virol 41: 462–477
Weber PC, Challberg MD, Nelson NJ, Levine M, Glorioso JC (1988) Inversion events in the HSV-1 genome are directly mediated by the viral DNA replication machinery and lack sequence specificity. Cell 54: 369–381
Weller SK, Aschman DP, Sacks WR, Coen DM, Schaffer PA (1983) Genetic analysis of temperature-sensitive mutants of HSV-1: the combined use of complementation and physical mapping for cistron assignment. Virology 130: 290–305
Weston K (1988) An enhancer element in the short unique region of human cytomegalovirus regulates the production of a group of abundant immediate early transcripts. Virology 162: 406–416
Weston K, Barrell BG (1986) Sequence of the short unique region, short repeats and part of the long repeat of human cytomegalovirus. J Mol Biol 192: 177–208
Whitton JL, Clements JB (1984) The junctions between the repetitive and the short unique sequences of the herpes simplex virus genome are determined by the polypeptide-coding regions of two spliced immediate-early mRNAs. J Gen Virol 65: 451–466
Wilkinson GWG, Akrigg A, Greenaway PJ (1984) Transcription of the immediate early genes of human cytomegalovirus strain AD169. Virus Res 1: 101–116
Worrad DM, Caradonna S (1988) Identification of the coding sequence for herpes simplex virus uracil-DNA glycosylase. J. Virol. 62: 4774–4777
Wright DA, Staprans SI, Spector DH (1988) Four phosphoproteins with common amino termini are encoded by human cytomegalovirus AD169. J Virol 62: 331–340
Wu CA, Nelson NJ, McGeoch DJ, Challberg MD (1988) Identification of herpes simplex virus type 1 genes required for origin-dependent DNA synthesis. J Virol 62: 435–443
Yang-Feng TL, Barton DE, Thelander L, Lewis WH, Srinivasan PR, Francke U (1987) Ribonucleotide reductase M2 subunit sequences mapped to four different chromosomal sites in humans and mice: functional locus identified by its amplification in hydroxyurea-resistant cell-lines. Genomics 1: 77–86
Zhang CX, Decaussin G, de Turenne Tessier M, Daillie J, Ooka T (1987) Identification of an Epstein-Barr virus-specific deoxyribonuclease gene using complementary DNA. Nucleic Acids Res 15: 2707–2717